用于发现复杂动力系统数据中流形维数和坐标的自编码器

Kevin Zeng、Carlos E. P´erez De Jes´us、Andrew J. Fox 和 Michael D.Graham*化学与生物工程系，威斯康星大学麦迪逊分校，美国威斯康星州麦迪逊市 53706

arXiv:2305.01090v3 [cs.LG] 6 Dec 2023

（日期：2023年12月7日）

摘要

尽管物理学和工程学中的许多现象在形式上具有高维特征，但其长期动力学往往位于一个低维流形上。本文提出一种自编码器框架，将隐式正则化与内部线性层及 L_2 正则化（权重衰减）相结合，以自动估计数据集的内在维数，生成正交流形坐标系，并提供环境空间与流形空间之间的映射函数，从而支持样本外投影。我们通过一系列来自复杂程度各异的动力系统的数据集，验证该框架估计流形维数的能力，并将其与其他最先进的估计器进行比较。我们分析网络的训练动力学，以深入了解低秩学习的机制；结果发现，各个隐式正则化层共同作用，不断强化低秩表示，甚至会在训练过程中进行自我修正。对该架构在线性情形下的梯度下降动力学进行分析后，我们揭示了内部线性层如何使包含所有层的“集体权重变量”更快衰减，以及权重衰减如何打破退化，从而推动系统沿着若无权重衰减便不会发生衰减的方向收敛。我们还展示了，该框架可通过仅使用流形坐标，为时空混沌偏微分方程生成数据驱动的动力学模型，从而自然扩展到状态空间建模和预测的应用。最后，我们证明该框架对超参数的选择具有稳健性。

*电子邮箱：mdgraham@wisc.edu

I. 引言

非线性耗散偏微分方程（偏微分方程）广泛用于描述物理学和工程学中的各种现象，这些现象呈现复杂的非线性行为、非平衡动力学，乃至时空混沌。尽管偏微分方程的状态空间在形式上是无限维的，但已知或推测耗散系统的长期动力学会收缩到一个有限维的不变流形上，我们将其记为 $\mathcal{M}$。[1–3]。同样的思想也适用于高维常微分方程系统（或离散化偏微分方程），无论如何，任何待研究系统的数据都是有限维的，因此我们将考虑维数为 d_m 、嵌入于环境空间 $\mathbb{R}^{d_u}$中的流形，其中通常有 $d_m \ll d_u$。也就是说，为了准确描述流形及系统的底层动力学，只需要 d_m 个相互独立的坐标（至少在局部范围内）。一般而言，不存在维数为 d_m 的全局坐标表示，但惠特尼定理保证可以找到具有 嵌入维数 $d_e \leq 2d_m$ 的全局表示 [4]。或者，原则上可以构造由维数为 d_m 的重叠坐标图组成的图册，以提供局部 d_m维表示 [4, 5]。我们主要致力于学习最小的全局流形表示（尽管我们将展示如何将本研究扩展到局部表示），并考虑 $d_e = d_m$ 的情形。

通过分析系统数据，为这些系统获得最简流形坐标描述，这对状态空间辨识、降阶建模与控制、系统可解释性等多种动力学应用，以及分类等许多下游任务都十分理想。然而，估计数据集的内在维数并获得流形坐标变换，通常并非易事。若只能获取以系统高维环境空间表示的数据，挑战便在于：1）确定 d_m；2）构建一个描述 $\mathcal{M}$ 中各点的坐标系；3）获得映射函数$\mathcal{E}$：$\mathbb{R}^{d_u} \to \mathbb{R}^{d_m}$ 和 $\mathcal{D}$：$\mathbb{R}^{d_m} \to \mathbb{R}^{d_u}$。为识别和建模这些系统的内在核心动力学，我们旨在仅利用高维环境空间数据训练一个统一框架，以应对这三项挑战。

这三项挑　　这些挑战已通过广泛多样的方法得到应对，但　　e

需要强调的是，很少有研究在单一框架内同时应对全部三个挑战——通常仅尝试解决识别流形维数这一首要挑战。对于复杂动力系统而言，许多源于系统理论的方法依赖高精度分析以及对底层方程的掌握。例如，Yang 等人[6], Yang 和 Radons [7] 使用协变 Lyapunov 向量，通过监测系统的 Lyapunov 谱何时开始快速下降，估计了 Kuramoto-Sivashinsky 方程（KSE）的流形维数。KSE 是一个形式上无穷维、但具有有限维动力学的系统；他们针对一系列参数进行了估计。Ding 等人[8]验证了这些结果，并估计了包含 KSE 长时间动力学的不变流形维数，其定义域大小为 $L = 22$；所采用的是 Floquet 模态方法，并将其应用于系统中识别出的有组织不稳定周期轨道。这些方法要求对控制方程进行高精度求解，并获取非常具体的动力学数据（例如周期轨道）；对于 Navier-Stokes 方程等更复杂的系统而言，这些任务并非易事，甚至可能无法处理。此外，当控制方程未知，或数据来自一般时间序列而非精确规定的轨迹时，这些方法也不适用。因此，本文将不重点讨论这些方法。

为了采用更具普适性和数据驱动的方法，人们已在模式识别、信息科学和机器学习领域探索估计表示一组数据所需自由度数量的任务，以避免信息损失。这些方法通过对数据集进行全局或局部分析，得出 d_m 的估计值（或某个上界）。

全局方法通过多种方式应对这一挑战。线性全局投影方法，例如主成分分析（PCA）[9] 及其变体（如稀疏主成分分析[10] 和贝叶斯主成分分析 [11]），会确定一个线性子空间，使数据在该子空间上的投影误差最小。这些方法的优点在于，它们不仅计算上易于处理，还能直接提供映射到低维表示的映射函数。然而，由于它们是线性方法，通常会高估 d_m，因为要表示维数为 d_m 的曲流形上的数据，至少需要 $d_m + 1$个坐标。

非线性主成分分析，或一般的深度自编码器，通过执行自联想任务的神经网络处理非线性 [12]。自编码器可用于通过追踪来估计 d_m 通过追踪

均方重构误差（均方误差）关于神经网络瓶颈维度的函数 d_z。如果均方误差在某个阈值 d_h以上显著下降，则可以推断，表示系统数据所需的最小自由度数量已经达到。在复杂高维动力系统（包括离散化耗散型偏微分方程）的应用中，Linot 和 Graham[13, 14] 以及 Vlachas 等人 [15] 以这种方式使用欠完备自编码器来估计来自 KSE 的数据的流形维数。然而，随着系统复杂度和维数增加，均方误差的下降变得越来越不明显[13, 16–18]。此外，这类方法的一个实际缺点是，需要训练一系列具有不同 d_z的独立神经网络。

为了构建更加自动化的基于自编码器的框架，一些研究采用启发式假近邻算法（FNN）[19]，以确定来自具有 $d_m = 3$流形的系统的时间序列信号在状态空间重构中的嵌入维数。具体而言，Gilpin [20] 引入了基于假近邻度量的额外损失，以惩罚编码器输出。然而，这种公式化既惩罚冗余潜变量，也惩罚用于表征流形的潜变量，因此对正则化项 [21]高度敏感。为解决这一问题，Wang 和 Guet [21] 引入了注意力图，根据假近邻度量显式屏蔽多余的潜变量。实际上，这些框架需要在每次训练迭代中，针对一系列嵌入维数反复计算训练数据点之间的欧氏距离；对于复杂度和维数不断增加的系统而言，这并不理想。此外，假近邻算法的目标是确定嵌入维数，而嵌入维数通常高于流形维数。

若干值得关注的降维方法采用局部计算。这些方法中有很大一部分属于称为多维尺度分析（MDS）的方法类别，其重点是保留数据的某些局部或成对特征。这些方法包括拉普拉斯特征映射 [22], t 分布随机邻域嵌入 [23], ISOMAP [24] 和局部线性嵌入 [25]。然而，这些方法与本文目标之间的一个主要区别在于，它们需要预先选择一个流形维数来嵌入数据，并且通常用于数据可视化。ISOMAP [24], 虽能通过误差曲线给出对 d_m 的“目测”估计，但难以处理维数较高的数据 [26]。另有一些基于原理的 d_m 估计方法，例如 Levina-Bickel 方法 [27] 以及

Little-Jung-Maggioni 方法（多尺度SVD） [28], 通过对数据点邻域所得的估计值取平均来估计 d_m 。下文将进一步讨论多尺度SVD和 Levina-Bickel 方法。重要的是，所有这些局部方法都缺少以下一项或多项能力：估计 d_m、将新的样本外数据点投影到流形坐标中，或为 d_m维表示提供一个坐标系。

在本研究中，我们采用一种深度自编码器框架来应对上述三个挑战，该框架会将潜在表示中数据协方差的秩降至最低。此秩将等于数据所处流形的维数 d_m 。我们的框架利用两种驱动秩降低的机制。第一种称为隐式正则化，这是在深度线性网络（即多个线性层串联）中进行基于梯度的优化时观察到的一种现象，可产生低秩解 [29]。尽管一系列线性层在功能和表达能力上与单个线性层完全相同，但二者的学习动态有所不同。该现象的作用机制仍是一个持续研究的领域，研究重点主要在矩阵 [29, 30] 和张量分解 [31]。值得注意的是，研究发现，对于随机满秩噪声 [30], 等无结构数据集，不会出现隐式正则化，这表明该现象取决于数据的内在结构。近期，Jing 等人 [32], 将隐式正则化推广至自编码器（隐式秩最小化自编码器，IRMAE），以学习低秩表示，从而改进基于图像的分类和生成问题中的表示学习；他们的观察构成了本研究的基础。

第二种低秩驱动力是 L_2 正则化，通常根据它与梯度下降结合时的作用称为：权重衰减。权重衰减是深度学习中一种常用的权重正则化机制，它迫使网络在学习问题的标准损失 L与网络权重的属性 θ之间做出权衡，

$$\mathcal{L} = L + \frac{\lambda}{2}\|\theta\|_p^2. \tag{1}$$

最近，Mousavi-Hosseini 等人 [33] 证明，在两层神经网络中，只有在线随机梯度下降（随机梯度下降）与权重衰减相结合时，第一层权重才会收敛到由目标函数张成的最小主子空间。作者发现，权重衰减使随机梯度下降能够避开主子空间之外的临界点。此处

我们展示了将权重衰减与隐式正则化的自编码器相结合时所获得的类似协同结果。

本研究旨在证明，在深度自编码器中结合隐式正则化与权重衰减的方法——我们称之为带权重衰减的隐式秩最小化自编码器，简称*IRMAE-WD*——可应用于位于维数为 $d_m < d_u$的流形上的数据集，以实现以下目标：1）估计数据所在流形的维数；2）获得描述该流形的坐标系；3）获得往返于流形坐标与原始空间的映射函数。值得强调的是，IRMAE-WD 在构造上会生成按方差排序的正交流形坐标基，且不依赖对神经网络进行大量参数扫描[13, 15] 或使用外部估计器 [20, 21] ——只需对流形维数作出一个合理的上界估计即可。（如果这一估计不合理，分析结果也会表明这一点。）这些特性使 IRMAE-WD 框架成为数据驱动降阶／状态空间建模及许多其他下游任务的理想起点。

本文其余部分安排如下：第 II 节介绍 IRMAE-WD 框架。第 IIIA 节将该框架应用于一系列数据集，涵盖从合成数据集到呈现复杂混沌动力学的物理系统，包括洛伦兹系统、Kuramoto-Sivashinsky 方程和 λ-ω反应扩散系统。第 IIIB 节概述模型性能对超参数的敏感性。第 III C 节将该框架估计复杂数据集底层维数的能力，与若干最先进的估计器进行比较。第 IIID节展示如何自然地扩展该框架，以用于流形坐标中的状态空间建模和动力学预测等下游任务。最后，第 IIIE节考察 IRMAE-WD 的训练动力学，以探明“空间”（即数据表示在经过架构时如何变换）和“时间”（即数据表示在训练过程中如何变换）两个方面低秩特性的成因。我们由此深入理解网络学习，并通过分析线性自编码器这一特殊情况，阐释隐式正则化和权重衰减如何产生协同效应。附录 A 总结了我们的架构，附录 B 详细介绍了在 MNIST 手写数字数据集上的应用，附录 C 则包含对线性自编码器的分析。

II. 公式化

我们提出的框架采用自编码器架构。自编码器由两个子网络组成，即编码器和解码器，两者通过潜在隐藏层相连。对于降维问题，该潜在隐藏层通常是一个限制尺寸的瓶颈，会明确限制用于表示输入数据的自由度数量。这种架构称为标准自编码器，它迫使编码器网络 $z = \mathcal{E}(u;\ \theta_E)$ 将输入数据 $u \in \mathbb{R}^{d_u}$ 压缩为紧凑表示 $z \in \mathbb{R}^{d_z}$，其中 $d_z < d_u$。解码器 $\tilde{u} = \mathcal{D}(z;\ \theta_D)$ 则执行相反的任务，学习从压缩表示 z 重构输入 $\tilde{u} \in \mathbb{R}^{d_u}$。自编码器通过最小化均方误差（MSE）或重构损失进行训练

$$\mathcal{L}(u;\theta_E,\theta_D) = \langle ||u - \mathcal{D}(\mathcal{E}(u;\theta_E);\theta_D)||_2^2 \rangle \tag{2}$$

此处 $\langle\cdot\rangle$ 是对一个训练批次的平均值，而 θ_i 对应于每个子网络的权重。随后，我们通过在编码器网络与解码器之间添加一个额外的线性网络 $\mathcal{W}(\cdot;\ \theta_W)$，偏离了标准自编码器架构：即 $z = \mathcal{W}(\mathcal{E}(u;\ \theta_E);\ \theta_W)$，其中 $\mathcal{W}(\cdot;\ \theta_W)$ 由 n 可训练的线性权重矩阵构成，记为 W_j（即线性层），其大小为 $d_z \times d_z$，并按串联方式排列，这与 Jing 等人[32]的做法相同。尽管 $\mathcal{W}$与标准自编码器相比增加了额外的可训练参数，但并未赋予神经网络额外的表达能力，因为串联的线性层与单个线性层具有相同的表达能力。因此，这两个神经网络的有效容量相同。重要的是，我们使用图 1a 所示的权重衰减来训练该框架，并采用自联想任务，

$$\mathcal{L}(u;\theta_E,\theta_W,\theta_D) = \langle ||u - \mathcal{D}(\mathcal{W}(\mathcal{E}(u;\theta_E);\theta_W);\theta_D)||_2^2 \rangle + \frac{\lambda}{2}\|\theta\|_2^2. \tag{3}$$

在此，我们从两个方面将 IRMAE-WD 与旨在寻找最小或低维表示的典型自编码器进行对比。首先，我们不采用标准自编码器通常使用的参数化扫描d_z，而是猜测一个单一的 $d_z > d_m$，并依靠隐式正则化和权重衰减，使潜在空间趋向于近似的最小秩表示。如果发现该秩等于 d_z，则可以增大 d_z 并重复分析。

正则化网络训练完成后，我们对潜在数据矩阵的协方差矩阵进行奇异值分解 Z （即编码数据矩阵），以获得奇异向量矩阵 U 和奇异值 S，如图1b所示。该谱中显著奇异值的数量可用于估计 d_m，因为每个显著值都代表在潜在空间中表示原始数据所必需的一个坐标。（更准确地说，我们得到的是 d_e的估计，不过如下文所示，可以对数据子集进行分析，以便在 d_m 的情况下确定 $d_m < d_e$。）

如图 1c所示，我们可以自然地将 z投影到 U^T上，从而得到$U^T z = h^+ \in \mathbb{R}^{d_z}$，其中 h^+的每个坐标彼此正交，并按贡献大小排序。由于 $UU^T = I$，我们可以通过将 z投影到 $\tilde{z}$上，重构 h^+，即 U。重要的是，该框架会自动发现一个潜在空间，其中编码后的数据仅跨越 d_m （这反映在显著奇异值的数量上）；数据只占据潜在空间中与这些显著奇异值对应的奇异向量所指向的方向。换言之，编码后的数据不会跨越对应奇异值近似为零的奇异向量所指向的方向，并且 $UU^T z \approx \hat{U}\hat{U}^T z$成立，其中 $\hat{U}$ 是截断后的奇异向量，仅包含对应奇异值不近似为零的奇异向量。

这一观察使我们能够通过简单地将 z投影到 $\hat{U}$上，得到最小表示 $\hat{U}^T z = h \in \mathbb{R}^{d_m}$，从而分离出一个最小、正交的坐标系。我们将其称为流形表示，如图 1d所示。由于 $UU^T z \approx \hat{U}\hat{U}^T z$和$u \approx \mathcal{D}(\mathcal{W}(\mathcal{E}(u;\ \theta_E);\ \theta_W);\ \theta_D)$，我们可以以最小损失从流形表示 h变换到环境表示 u。

为了以可处理的方式深入了解带有隐式正则化和权重衰减的自编码器的学习机制，我们在附录 C 中分析了梯度下降对一个线性自编码器的动力学。该自编码器作用于协方差矩阵为对角矩阵、秩为$r(= d_m$的数据。在这种情况下，权重矩阵存在一个解族，其中每个矩阵的秩均为 r。分析表明，存在一种“集体”衰减模式，趋向低秩解族，且所有权重矩阵彼此耦合。该模式的衰减率和均方误差的衰减率均按 $2 + n$缩放，其中 n 是内部（方形）线性层的数量。没有权重衰减时，存在特征值为零且在训练过程中不会衰减的方向。加入权重衰减后，这些原本为零的

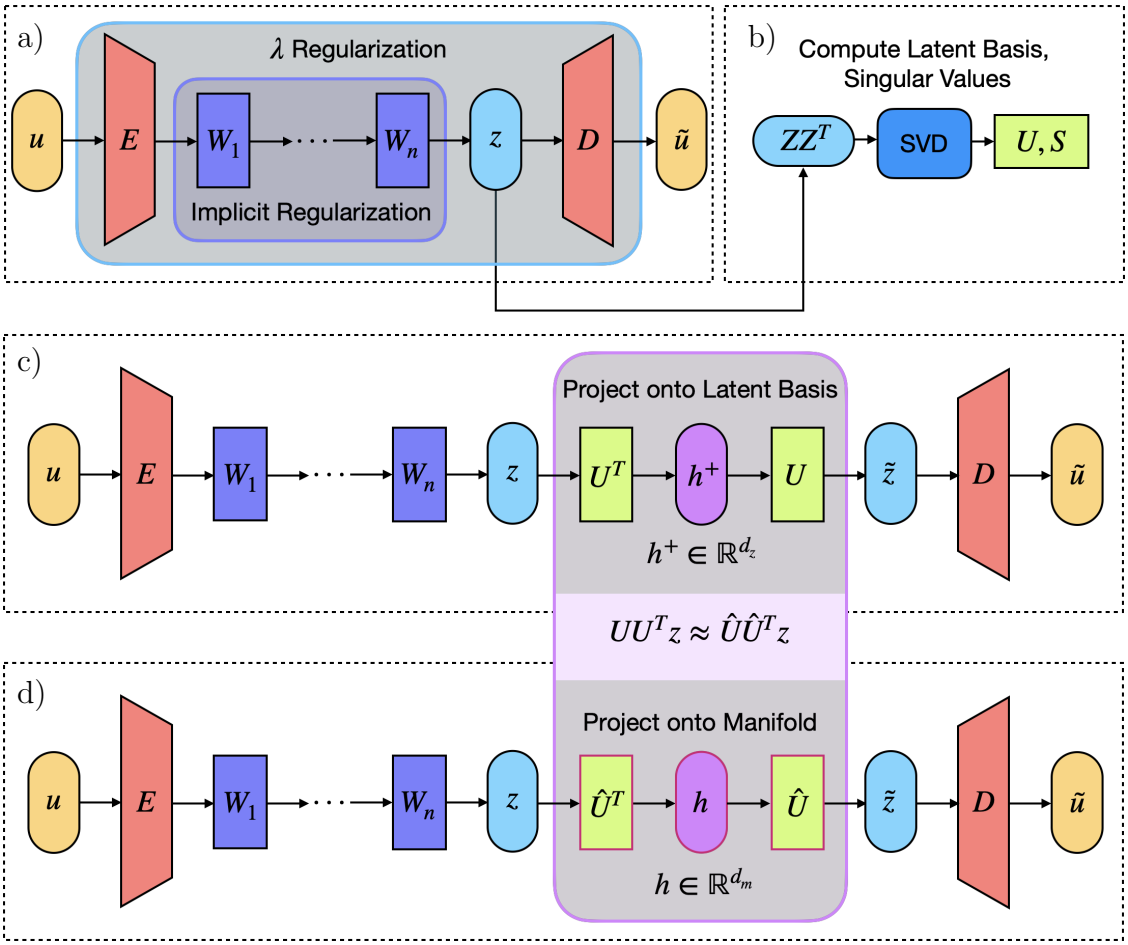

图1：我们的隐式 λ 权重衰减正则化深度自编码器框架：a）带有正则化机制的网络架构，b）所学习的潜在数据表示 Z的协方差的奇异值分解，c）将潜变量投影到流形坐标上，d）将潜变量单独投影到流形坐标上。

特征值变为负值，使所有方向上的状态都能衰减至低秩解。在第 IIIE节中，我们通过实证观察到，非线性网络中线性层的梯度更新和权重矩阵呈现出向低秩解收敛的趋势。

应用本方法时，一个重要的实践细节是为随机梯度下降过程选择优化器。我们发现，使用 AdamW 优化器 [34] 而非标准 Adam 优化器至关重要。之所以要区分二者，是因为在常用的 Adam 优化器中直接应用权重衰减（L_2 正则化），会导致梯度幅值较大的权重受到不成比例的正则化 [34]。AdamW 将权重衰减与自适应梯度更新解耦。我们发现，

使用基础 Adam 优化器并进行 L_2 正则化，可能会导致对参数高度敏感，并产生虚假结果。

III. 结果

A. 流形维数估计：示例系统

现在，我们研究将 IRMAE-WD 应用于一组复杂度不断增加的数据集，范围从嵌入有限维环境空间中的线性流形，到嵌入形式上无限维环境空间中的非线性流形。

1. 线性嵌入有限维环境空间中的数据

我们首先以一个简单的数据集作为基准，评估 IRMAE-WD。该数据集由线性嵌入到 20 维环境空间中的 5 维噪声组成。由于该数据集恰好张成 5 个正交方向，并且采用线性嵌入，因此主成分分析（PCA）能够从数据中提取 d_m ，这可以通过数据矩阵协方差的奇异值谱加以识别。如图 2a 所示，展示了通过 PCA 获得的奇异值、IRMAE 在无权重衰减和有权重衰减情况下学习到的潜变量的奇异值，以及在架构上与 IRMAE-WD 相同但不进行任何正则化（即没有 W 和 $\lambda = 0$）的标准自编码器的奇异值。对于标准自编码器，虽然奇异值 σ_i 在索引 $i > 5$处略有下降，但其谱较为宽广且衰减缓慢，这表明所学习的潜在表示本质上是满秩的。换言之，当瓶颈层具有过剩容量时，标准自编码器会利用所有可用的潜变量。相比之下，对于 IRMAE-WD， $i > 5$ 的奇异值降至 $\sim 10^{-16}$，与 PCA 的情况相同。这表明 IRMAE-WD 能够自动学习一种表示，分离出表示数据所需的最小维数。

我们在此进一步强调两个重要观察：1）仅使用权重衰减的自编码器不足以学习稀疏表示——其表现与标准自编码器非常相似；2）仅使用隐式正则化的自编码器（如 Jing 等人 [32], 所采用的方法）会使 $i > 5$ 的 σ_i 急剧下降，但与同时采用线性层和权重衰减相比，效果并没有那么显著。本文将探讨这一现象

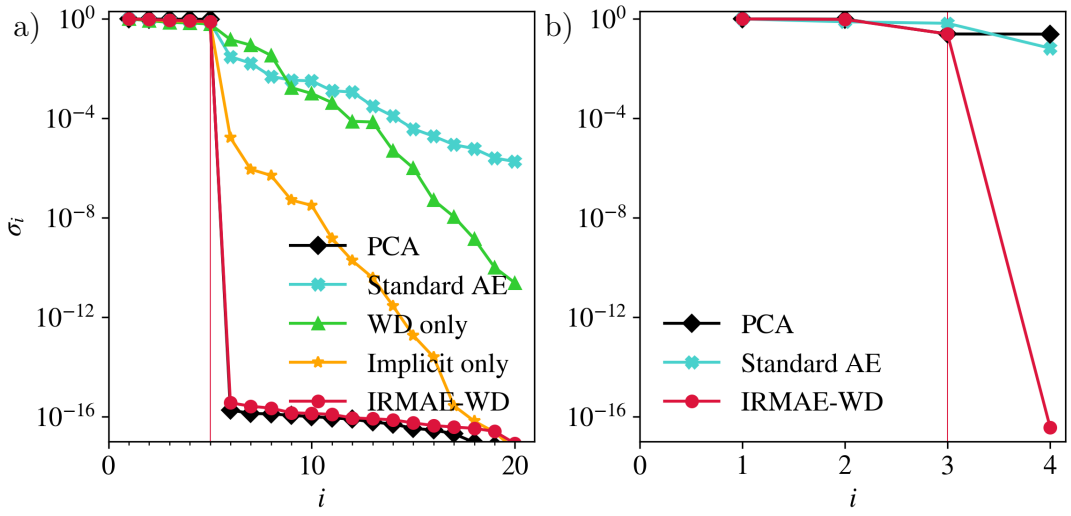

图2：应用多种自编码器方法得到的潜在数据协方差的归一化奇异值 σ_i，这些方法分别用于嵌入 $\mathbb{R}^{20}$ 的五维线性流形（a）和嵌入$\mathbb{R}^4$的三维非线性流形（b）。同时给出了主成分分析和未进行正则化的标准自编码器得到的谱。d_m 的值以红色竖直参考线标出。

在第 IIIE 节和附录 C中。

2. 非线性嵌入有限维系统：阿基米德螺线 Loren z

现在，我们将关注具有非线性嵌入流形的非线性有限维动力系统数据。具体而言，我们采用 Lorenz ‘63系统 [35],

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z \end{aligned} \tag{4}$$

该系统在 $\mathbb{R}^3$ 中呈现混沌动力学，并通过以下映射将数据集缠绕在阿基米德螺线上，从而将此系统非线性地嵌入 $\mathbb{R}^4$：

$$[x, y, \alpha z \cos(\alpha z), \alpha z \sin(\alpha z)] \to [u_1, u_2, u_3, u_4],$$

与 $\alpha = 0.2$。

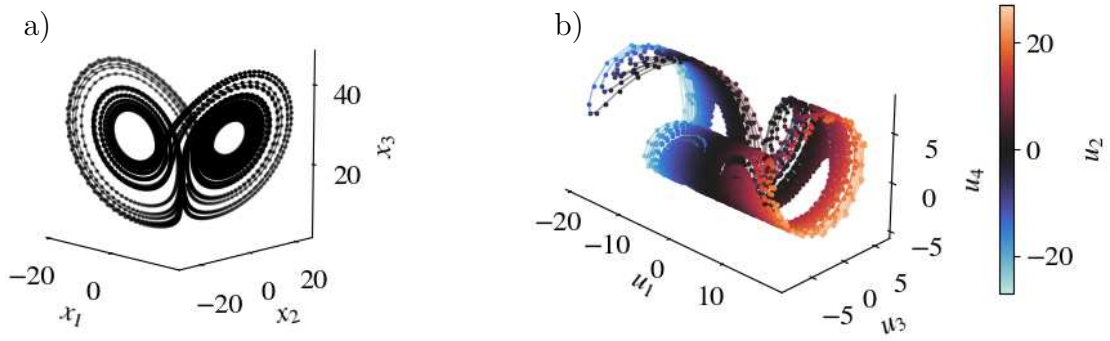

图 3：a) 三维洛伦兹'63方程和 b) 四维阿基米德洛伦兹方程的动力学。颜色对应嵌入中的变量 u_2，而空间坐标对应 u_1、u_3 和 u_4。

对于参数 $\sigma = 10$，$\rho = 28$，$\beta = 8/3$，Lorenz '63 呈现出混沌动力学。换言之，该系统的底层动力学位于一个非线性三维流形上，而该流形非线性嵌入在四维环境空间中。如图2b所示，我们表明，IRMAE-WD 正确地判定该系统可由 3 个潜变量进行最小表示。相比之下，主成分分析未能识别数据的底层结构。在这里，主成分分析谱无法正确估计 d_m，因为线性方法本质上无法以最小表示捕捉流形的非线性／曲率。最后，采用 $d_z > d_m$ 的标准自编码器也无法自动学习最小表示，因为它在潜在空间中得到的是满秩数据协方差。

3. 全局流形估计与局部估计：二维环面上的拟周期动力学

现在，我们将目光转向一条 $\mathbb{R}^3$ 轨迹：它以极向和环向速度穿过二维环面的曲面，由此产生拟周期动力学，如图4a所示。经过无限长的时间后，粒子将稠密地覆盖环面曲面。尽管该系统位于二维流形上，但这个流形的拓扑并不平凡，因此无法获得单一的全局表示[5]。在此，我们将 IRMAE-WD 应用于该数据集，其中包含沿轨迹采集的三个坐标的快照。在这个例子中，我们还采用了过完备的网络设计，$z \in \mathbb{R}^{10}$以此突出说明，即使$d_z > d_u$，多余的自由度仍能被正确消除。我们在图5a 中展示

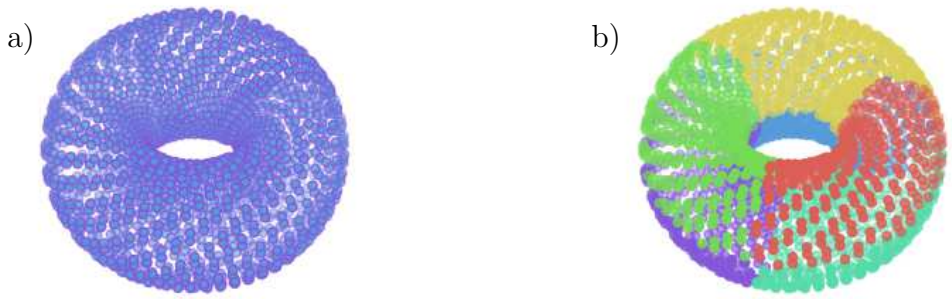

图4：环面上的拟周期动力学：a) 全局，b) 局部补丁。

当 IRMAE-WD 通过整个数据集进行训练，以学习全局表示时，它（正确地）获得了一个三维潜在空间——该流形的嵌入维数为 $d_e = 3$。然而，正如 Floryan 和 Graham 所述 [5], ，将流形分解为由重叠坐标图构成的图册，可以捕捉包含这些数据的流形的内在维数。在图5b中，我们展示了 IRMAE-WD 应用于同一数据集的结果；该数据集先被划分为若干补丁，这些补丁通过 k均值聚类确定，如图4b所示。结果表明，对于每个子域，IRMAE-WD 都能自动学习数据的最小二维表示，同时舍弃其余多余的自由度。通过这种方式，可以将 IRMAE-WD 应用于数据局部区域，以估计其内在维数。

4. “无限维”系统中的非线性流形：Kuramoto-Sivashinsky 方程和反应-扩散系统

现在我们将注意力转向两个形式上具有无限维数的耗散非线性系统——非线性偏微分方程，它们经过离散化后成为高维常微分方程系统。首先，我们研究一维 Kuramoto-Sivashinsky方程（KSE）：

$$\frac{\partial v}{\partial t} = -v\frac{\partial v}{\partial x} - \frac{\partial^2 v}{\partial x^2} - \frac{\partial^4 v}{\partial x^4} \tag{5}$$

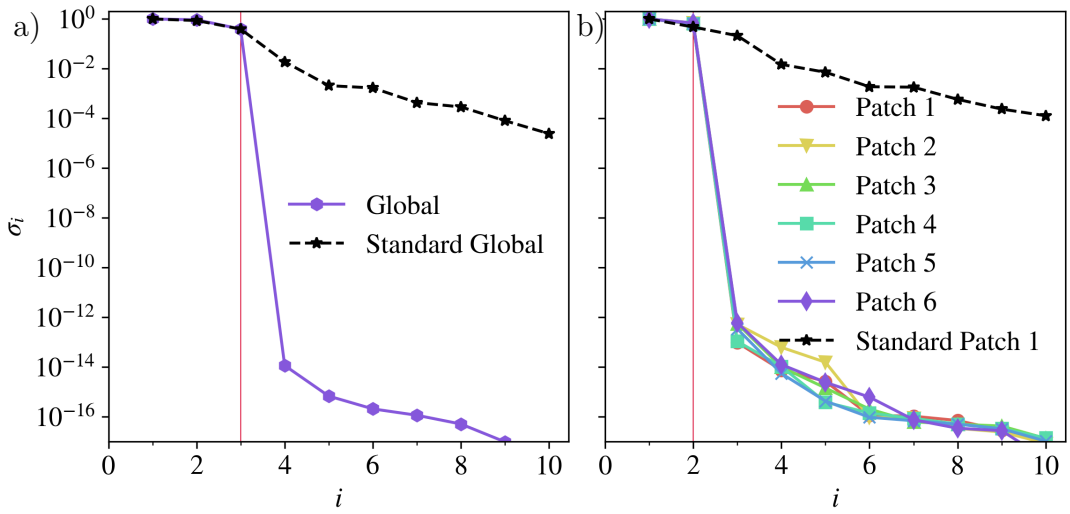

图 5：应用于 a) 全局环面数据集和 b) 环面数据集的局部补丁的 IRMAE-WD 所学习潜在空间的归一化奇异值 σ_i。标准自编码器潜在空间的结果以黑色显示。 d_m 的值由红色竖直参考线标出。

在长度为 L 的区域内，并采用周期性边界条件。对于较大的 L，该系统呈现出丰富的时空混沌动力学，因此已成为复杂非线性系统研究中的常见测试案例。为了对这一形式上“无限维”的系统进行分析，状态快照将由该区域内等距网格点上采样的解值构成。我们应用 IRMAE-WD 提取不同区域大小下动力学所对应的底层流形维数，首先重点研究 $L = 22$，该系统呈现时空混沌动力学，且已得到广泛研究。该系统的一条示例轨迹如图6a所示。尽管该系统在形式上是无限维的，但其动力学实际上由一个非线性嵌入的 8 维流形所支配，这一点已由多种方法证实[6, 8, 13, 36]。我们使用由 40,000 个快照组成的数据集，这些快照在 64 个网格点上采样，并选择瓶颈层维数为 $d_z = 20$，在图7a 中展示了 IR MAE-WD 得到的奇异值在索引超过 8 后急剧下降，这表明我们以自动且直接的方式学习到了一个维数为 $d_m = 8$的潜在空间。相比之下，无论是主成分分析还是标准自编码器，在整个索引范围内都未能使奇异值出现显著下降。

随着区域尺寸增大， LKSE 的时空动力学变得更加复杂。图6b-d 展示了 $L = 44$、66 和 88 时动力学的时空图，分别在

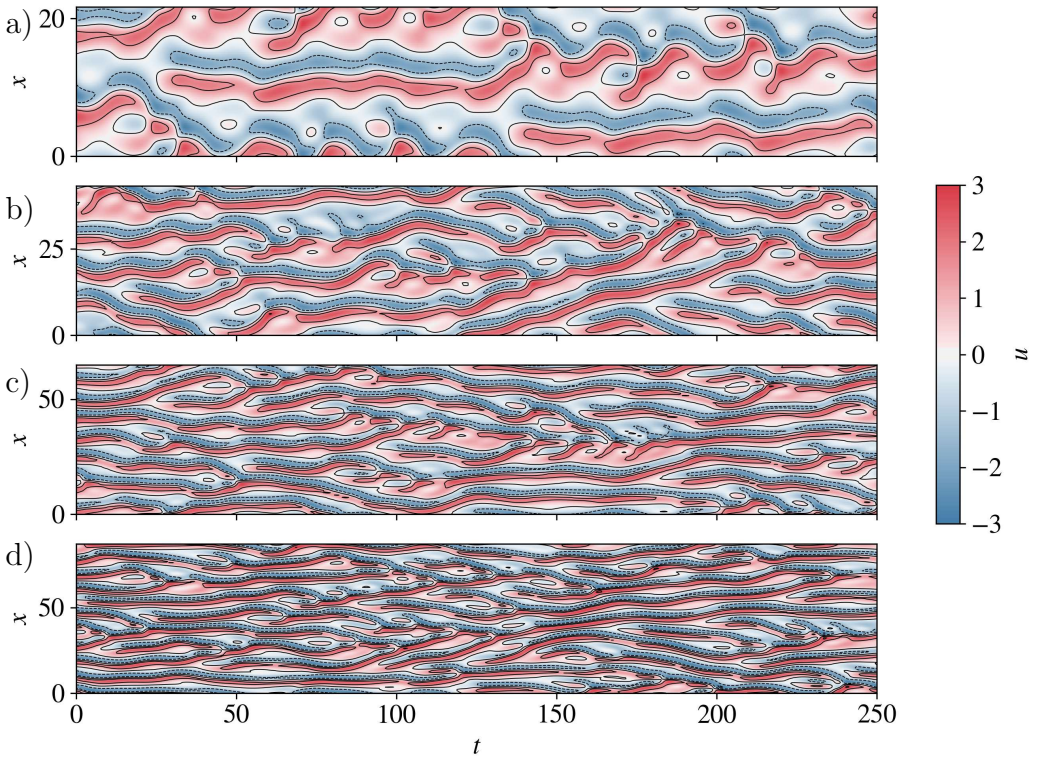

图 6：典型的演化　KSE 在域尺寸 a) $L = 22$、b) $L = 44$、c) $L = 66$ 和 d) $L = 88$ 下的解　　.

分别由包含 64、64 和 128 个点的均匀空间网格采样。图7b-c 展示了 $L = 44$ 和 66 的潜在空间协方差的奇异值谱，同样显示在明确定义的索引值处下降了 > 10个数量级，表明流形维数分别为 $d_m = 18$和 28。这里我们强调，先前的自编码器方法[13, 14], 利用均方误差随 d_z 变化的趋势来估计 d_m，难以区分这些区域尺寸下的流形维数，而 IRMAE-WD 则能给出明确的数值。以往基于详细且复杂的轨迹分析对动力学进行高精度分析的研究表明，KSE 的流形维数随区域长度 L [6, 36]线性增长。图7d 展示了我们利用 IRMAE-WD 得到的 d_m 与 L 之间的变化趋势：无需获取底层控制方程或周期解，我们就能直接恢复这种线性增长关系。

第二个无　　限维系统是 λ-ω反应　　-

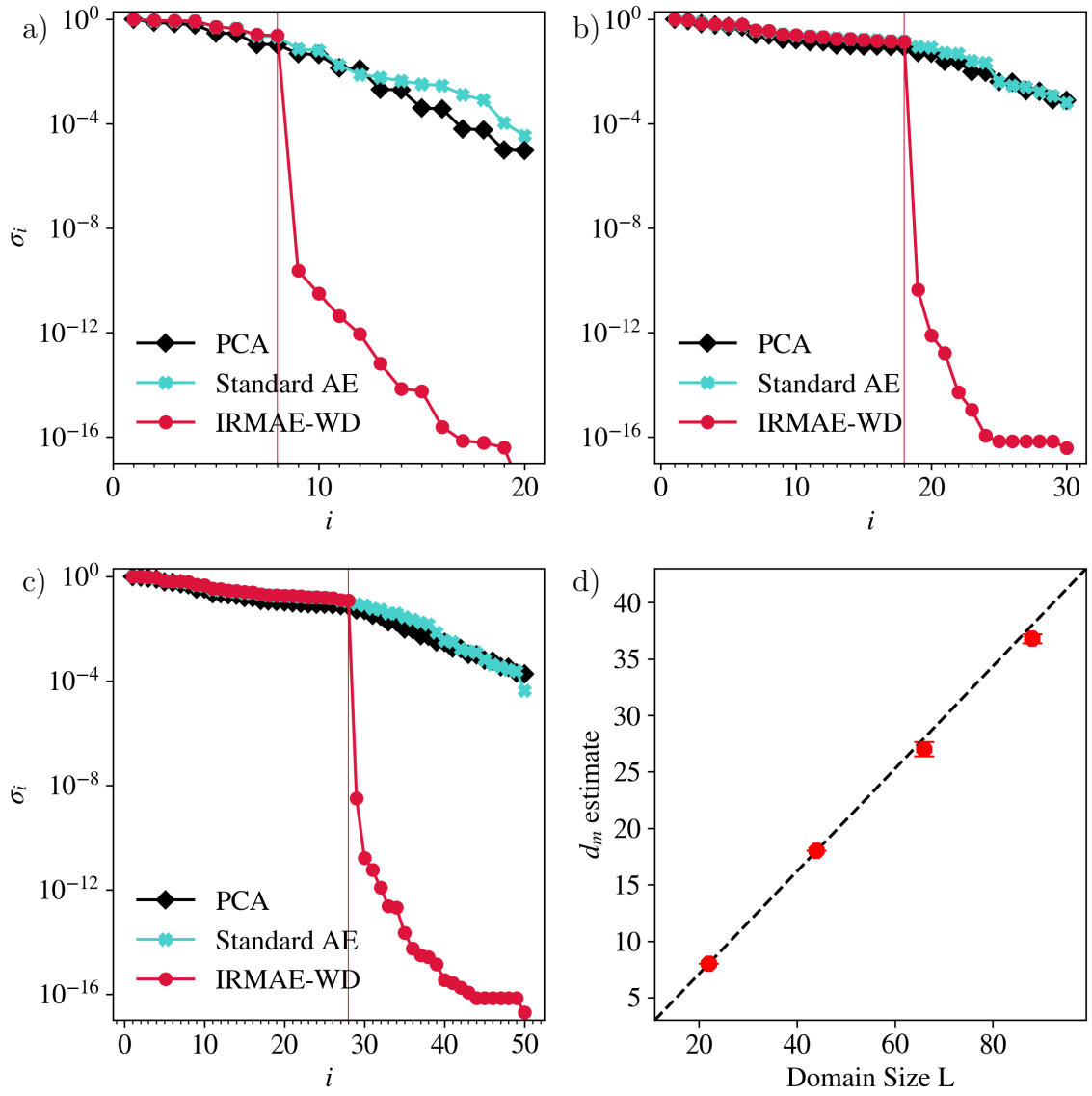

图7：KSE 的 IRMAE-WD 学得的潜在空间的奇异值 σ_i：a) $L = 22$、b) $L = 44$、c)$L = 66$，以及 d) 对 d_m 的估计。该估计为 5 个随机初始化模型的平均值，以 L 为函数变量，标准差用误差棒表示。a)-c) 中还显示了主成分分析和未进行正则化的标准自编码器得到的谱，并用红色竖直参考线标出 d_m 的值。

由以下方程支配的二维空间扩散系统

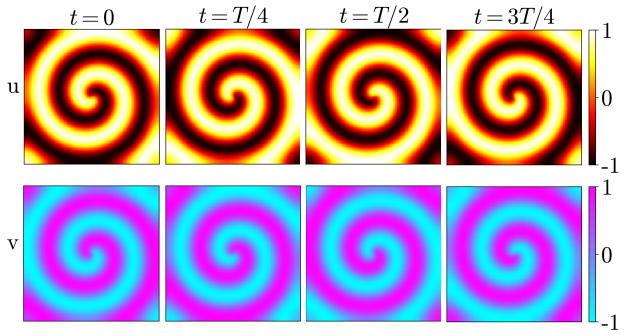

图 8：lambda-omega 反应扩散系统产生的螺旋波的一个周期 T。

$$\begin{aligned}\frac{\partial u}{\partial t} &= \left[1-\left(u^2+v^2\right)\right]u+\beta\left(u^2+v^2\right)v+d_1\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)\\ \frac{\partial v}{\partial t} &= -\beta\left(u^2+v^2\right)u+\left[1-\left(u^2+v^2\right)\right]v+d_2\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)\end{aligned} \tag{6}$$

其中 $d_1 = d_2 = 0.1$，且 $\beta = 1$ 对于 $-10 \leq x \leq 10$，$-10 \leq y \leq 10$。该系统此前已在 [5, 37, 38]中得到研究。该系统的长时间动力学会塌缩到状态空间中的一个吸引极限环上，其形式为螺旋波；通过单一的全局表示 [5]，即可在二维潜在空间中对其进行完整描述。我们将使用从 101×101 网格中等距网格点的解值采样得到的状态快照来分析该系统，从而得到环境维数 $\mathbb{R}^{20402}$。我们生成了一个包含 201 个快照的数据集，快照之间均匀间隔 0.05 个时间单位，覆盖略多于螺旋波一个周期的时间；螺旋波的一个周期如图8所示。我们采用瓶颈层维数为 $d_z = 10$的 IRMAE-WD，并在图9中展示了奇异值在索引 2 以上急剧下降，这表明我们已经自动且直接地学习到了维数为 $d_m = 2$的潜在空间。

B. 鲁棒性与参数敏感性

在下一节中，我们概述 IRMAE-WD 的参数鲁棒性，重点关注 KSE $L = 22$ 数据集。我们选择该数据集，是因为它来自一个高维非线性系统，该系统由非线性流形上的动力学支配，而且具有相当

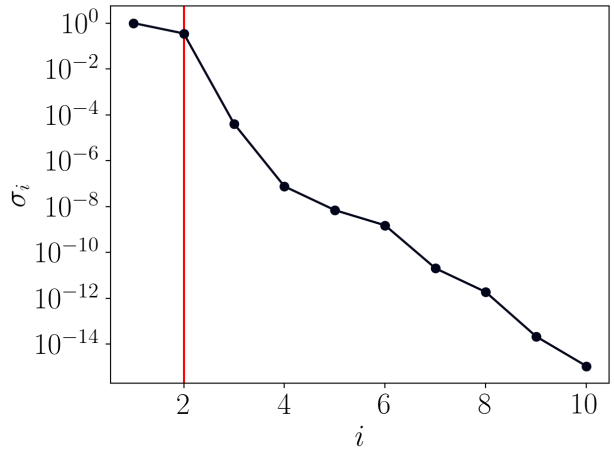

图9：λ-ω反应扩散系统中 IRMAE-WD 学得的潜在空间协方差的奇异值 σ_i 。

比典型的基准系统更复杂。

我们首先考察 d_m估计的准确性。根据多个来源一致的结果，正确值取为 $d_m = 8$。图10a展示了维数估计随线性层数 n 和权重衰减参数 λ变化的情况，图中的底行对应于采用 L_2正则化的标准自编码器这一情形 $n = 0$ 。我们强调，在广泛的 n 和 λ 取值范围内，该框架都能准确估计 d_m。只有当 n 和 λ 两方面的正则化都很强时，该框架才开始失效。在线性层未产生隐式正则化时，自编码器完全无法预测 d_m 。图10b展示了相同参数扫描下的测试均方误差性能。该指标对参数选择也相对不敏感；在表示中有效自由度为 d_m 的正则化模型，其重构误差与标准自编码器（左下角）相当。最后，对于理想的正则化模型，索引大于 d_m 的奇异值为零，但实际情况并非如此。在图10c中，我们量化了表示中来自谱尾部（即索引大于 d_m）的奇异值所贡献的总方差占比：
$\sigma^+ = \sum_{i=d_{m+1}}^{d_z} \sigma_i / \sum_{j=1}^{d_z} \sigma_j$。我们在此强调，在广泛的 n 和 λ取值范围内，尾部奇异值贡献的总方差或能量约为 10^{-9} ；而未正则化模型则贡献了不可忽略的 10^{-1}。最后，我们指出，我们没有观察到.....的强烈依赖性

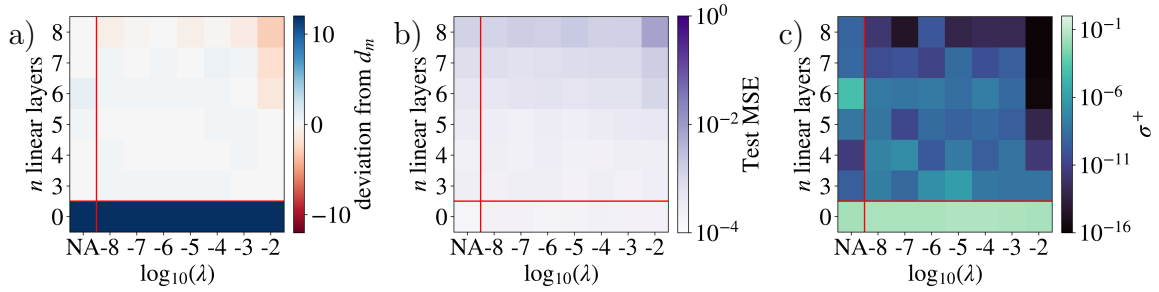

图10：对采用不同程度的隐式正则化和权重正则化、在 KSE $L = 22$ 数据集上训练的模型进行 n 和 λ 的参数扫描。颜色分别表示：a) 3 个模型相对于 d_m 的平均偏差，b) 3 个模型的平均测试均方误差，c) 3 个模型尾部奇异值所占比例的最低值（越低越好）。在最左侧标记为 NA 的列中，$\lambda = 0$。每个图的左下角对应一个标准自编码器。

只要 $d_z > d_m$，d_z 的选择不会影响结果。

C. 与其他方法的比较

在本节中，我们将 IRMAE-WD 与两种最先进的估计器进行比较：多尺度SVD（MSVD） [28]和 Levina-Bickel 方法 [27] ，因为这些方法旨在直接估计数据的流形维数。我们首先与 MSVD 方法进行比较，因为它完全由数据驱动，也依赖于分析数据的奇异值谱。MSVD 通过追踪从一组数据局部邻域中获得的奇异值谱的总体平均值，随邻域半径大小变化的情况，来估计 d_m。 r奇异值谱中间隙的出现r 会随着r 增大而发生，这与流形上的方向和由曲率导致的方向之间的分离相吻合。假定谱中的间隙表示流形维数，正如主成分分析对线性流形上的数据所做的那样。我们重新考察 KSE $L = 22$ 和 $L = 44$ 数据集，因为它们是具有非线性流形的非平凡复杂系统。在图11a 和图11b中，我们展示了将 MSVD 方法应用于这些数据集的结果。这里要强调的是，对于这些数据集，MSVD 无法明确识别 d_m；谱中并非只有一个间隙，而是存在多个间隙，如箭头所示。这很可能是 MSVD 的一项关键局限所致：它要求数据位于足够小的 r 局部邻域中，以便

准确地将高度非线性流形近似为平坦流形。也就是说，为了在极小邻域的极限下工作，MSVD 要求数据点集成中，在非常小的 r 范围内有足够数量的邻近点。在应用 MSVD 时，我们无法在不遇到缺少邻点的点云样本的情况下，使用较小的 r 值。许多复杂动力系统并未均匀地填充其底层流形，因此会形成高密度和低密度区域——事实上，混沌吸引子上的数据点呈分形分布，而非均匀分布。因此，很难收集到能以均匀密度表示流形的动力学数据；如果这些区域自然出现的概率较低，也很难收集足够的数据，使低概率区域达到较高密度。IRMAE-WD 不受这些局限的影响。

我们还将 Levina-Bickel 方法应用于这些相同的数据集。该方法利用最大似然框架，根据局部区域的数据估计其维数 [27]。在我们的应用中，我们按照 Levina 和 Bickel 的建议固定邻居数量[27], ，而不是固定邻域半径。对于我们的数据集，该方法也无法提供可靠的估计。我们在表I中总结了本节的发现。对于所考察的数据集，Levina-Bickel 方法似乎低估了维数，而 MSVD 的估计往往含糊不清。

表I：使用多种方法对 d_m 进行的估计。

数据集	d_m	多尺度SVD	Levina-Bickel	IRMAE-WD
Arch. Lorenz	3	2	2.09	3
KSE $L = 22$	8	6-8	3.99	8
KSE $L = 44$	18	8-20	7.00	18

D. 流形坐标中的降阶状态空间预测

如上所述，并如图1所示，将潜在空间数据 z 投影到其协方差矩阵的前 d_m 个奇异向量上，可得到流形表示 $h \in \mathbb{R}^{d_m}$。通过将编码和解码的定义简单扩展到 ，我们可以将环境空间中的数据快照映射到这种流形坐标表示：

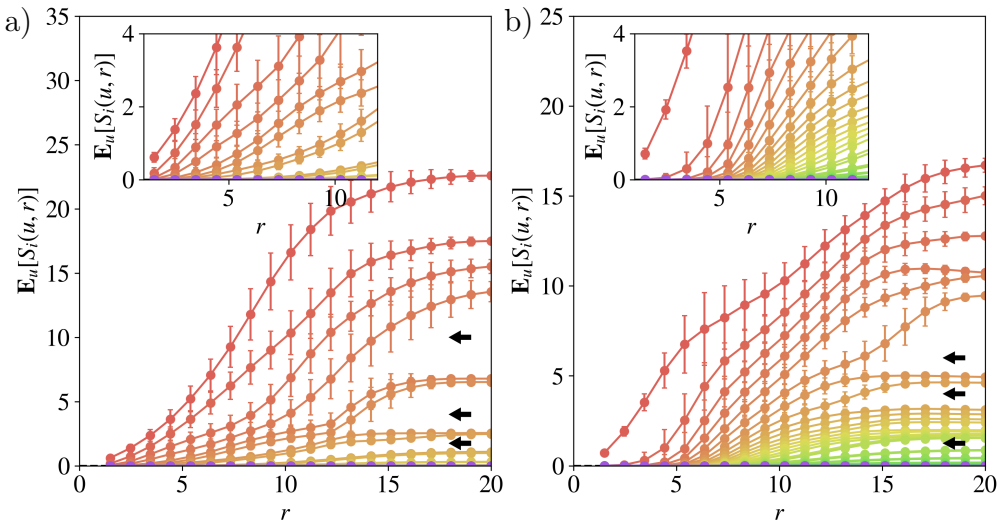

图 11：KSE 上的 a) $L = 22$ 数据集和 b) $L = 44$ 数据集的集成 MSVD 奇异值 S_i 随采样邻域半径 r 变化的情况。曲线的颜色对应谱的模态索引。箭头标出谱中出现的间隙，据此可以估计底层维数。

只需将编码和解码的定义简单扩展到 h:

$$
\begin{aligned}
h &:= \mathcal{E}_h(u; \theta_E, \theta_W, \hat{U}^T) = \hat{U}^T \mathcal{W}(\mathcal{E}(u; \theta_E); \theta_W) \\
\tilde{u} &:= \mathcal{D}_h(h; \theta_D, \hat{U}^T) = \mathcal{D}(\hat{U} h; \theta_D)
\end{aligned}
\tag{7}
$$

其中， $\mathcal{E}_h$ 和 $\mathcal{D}_h$ 只是概括了在 h 和 u之间进行映射所需的中间线性变换。通过扩展编码和解码的定义，我们现在已经 1）估计出 d_m，2）获得了用于参数化流形的坐标系 h ，以及 3）确定了在环境空间与数据流形之间来回映射的显式映射函数 $\mathcal{E}_h$ 和 $\mathcal{D}_h$ 。掌握这三者后，一个自然的应用便是状态空间建模与预测。我们在图12中展示了这一扩展的示意图；粉色的内部方框包含一个时间演化模块，用于将已转换为流形坐标表示 u_0 的初始条件 h_0 向前积分。既然已经明确确定了流形维数和坐标，就不再需要通过反复试验、测试不同维数的模型（例如[14]），来寻找最小维数的高保真时间演化模型。

继续之前的内容 .....例子之后，我们对这一方法以及

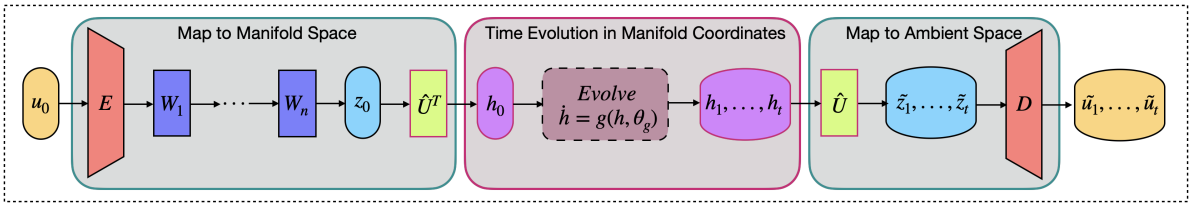

图 12：使用神经常微分方程（粉色部分）扩展 IRMAE-WD 框架，以在流形坐标系中进行预测的示意图。

本节讨论的问题作一些一般性说明。我们考虑的是长期动力学行为落在维数小于环境维数的不变流形上的确定性动力系统。一个简单的例子是具有稳定极限环的系统。从拓扑上看，这是一维的，但其嵌入维数为二。这意味着，要预测极限环上的动力学，必须且只需两个全局坐标，不多也不少。我们的目标是识别这些坐标以及其中的动力学。对于精确（流形）表示需要大量坐标的系统，仍有可能构建一个模型，用远少于真实不变流形维数的维数来预测系统的诸多方面，尤其是统计特性。这是一个常见目标，例如湍流的大涡模拟就是如此 [39]。但这并非我们在此试图实现的目标。

我们现在将上述 KSE 和反应扩散示例扩展到流形坐标中，以构建数据驱动的动力学模型。在这里，我们训练一个神经常微分方程 [40]，$\dot{h} = g(h; \theta_g)$，用于模拟 h 的时间演化，方法与 Linot 和 Graham [14] 所采用的方法相同。换言之，我们只需加入一个预测网络，使其在流形坐标表示中学习系统的动力学演化，h。图 13a 展示了 KSE 的一条示例轨迹。图 13b 展示了同一轨迹经过 $\mathcal{E}_h$ 编码后的流形表示。仅凭环境空间中的一个编码初始条件，我们便可在流形空间中对整个系统进行预测。图 13a 所用的同一初始条件所生成的预测轨迹见图 13d。自然地，通过 $\mathcal{D}_h$，可以完整地恢复该轨迹的环境表示，如图 13c 所示。比较上下两行可见，在流形坐标系中对时间演化的预测在近 50 个时间单位内具有定量准确性。需要强调的是，由于 KSE 是一个混沌动力系统，真实值与预测结果最终会发生偏离。

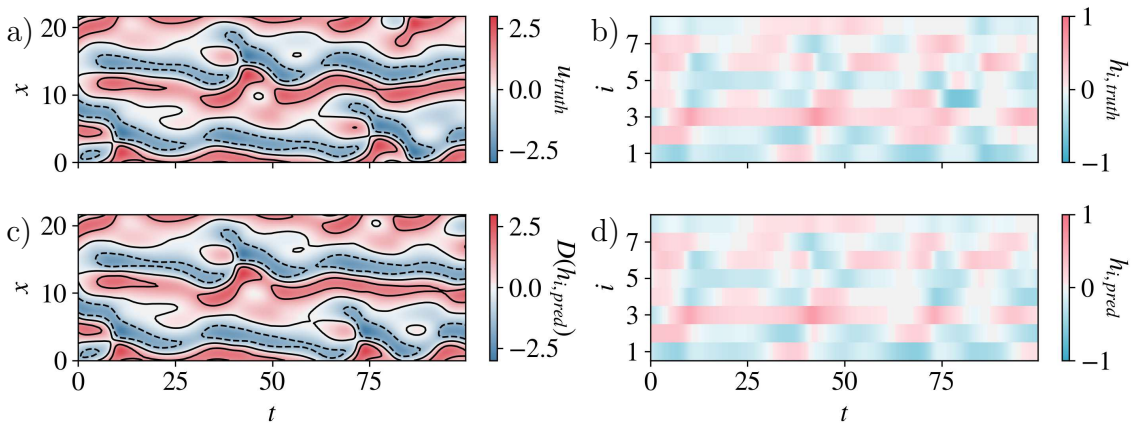

图13：KSE在a）环境空间中的真实轨迹，以及b）投影到学习得到的流形坐标表示上的真实轨迹示例。使用神经常微分方程在d）流形坐标中进行的时间序列预测，该预测从相同的初始条件开始。c）由神经常微分方程预测的流形轨迹解码得到的环境空间重构。

尽管如此，该系统的相关时间尺度（李雅普诺夫时间）为 ~ 20 个时间单位，而我们实现了约两个李雅普诺夫时间内的定量一致性。基于这一结果，我们指出，学习得到的流形坐标系适用于预测，而映射函数能够很好地重构环境空间。

我们进一步展示了 IRMAE-WD 利用 λ-ω反应扩散生成适用于高维环境系统的低维动力学模型的能力系统。我们训练了一个神经常微分方程 $\dot{h} = g(h;\ \theta_g)$，以在流形坐标 h 中对系统的动力学进行建模，方法与 Linot 和 Graham [14]所采用的方法相同。通过在环境空间中使用一个编码后的初始位置，我们可以在流形空间中预测完整环境系统的演化。图14a显示了一个周期后螺旋波的真实值，以及由神经常微分方程预测结果解码得到的相应环境空间重构。重构得到的螺旋波与真实值相匹配，因为预测的演化紧密跟踪了真实系统动力学的演化。在图14b中，我们展示了单个空间位置处真实值 u 和 v 分量的时间序列，以及根据预测轨迹得到的重构结果。预测的演化与真实值几乎完全一致，进一步证明了我们预测结果的定量准确性。

此外，我们简要指出，IRMAE-WD 与神经常微分方程在流形坐标中的演化相结合，已分别应用于微观结构预测

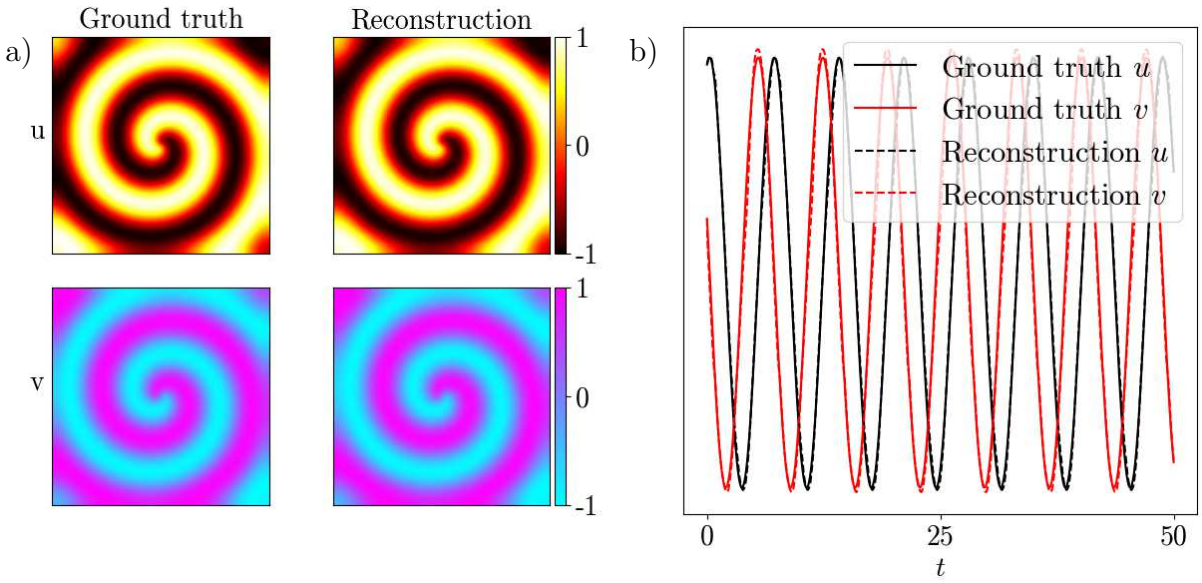

图14：a）λ-ω反应扩散系统产生的螺旋波经过一个周期后的真实值与预测结果；b） u 和 v分量在 $x = 1$, $y = 45$处的时间序列真实值与预测结果。真实值与预测结果均从相同的初始条件演化而来。

流动的复杂流体中的演化 [41]。在这种情况下，复杂流动中布朗刚性棒悬浮液的合成X射线散射图样数据被降维至五维流形，并学习了这些维数的演化，重构效果优异。

作为对流形坐标中时间演化建模讨论的总结，我们来探讨“端到端”建模是否可行或实用：这种建模方式能够从时间序列数据中同时确定流形维数、坐标和演化方程。尽管近期已有沿着这一思路提出的方法 [38], 但它们尚未应用于动力学复杂程度超过极限环的情形。对于环境维数较高且具有复杂混沌动力学的系统，轨迹需要多次遍历不变流形的每个区域，才能确定流形的真实形状。也就是说，必须先了解流形的全局“形状”，才能找到它的坐标表示。我们所知的所有流形维数估计方法都具有这一特点。利用从状态空间不同区域出发的轨迹集合，或许可以绕过这一问题（参见 [42]）；不过，这种方法超出了本文的范围。

E. 低秩表示学习的动力学

现在，我们将注意力转向理解如何自动学习近似最小表示。我们通过将网络视为一个动力系统来获得启示，其中“空间”对应网络中的层深度，“时间”对应训练轮次/迭代。通过这种方式，我们阐明了网络中的低秩行为“何时”以及“何处”出现。

更具体地说，我们将计算并跟踪一系列中间潜在表示、权重矩阵和更新梯度的奇异值谱，并分析它们随模型层和训练轮次的变化。我们将利用这些谱来估计秩（依据所考察矩阵的奇异值谱中显著间隙的位置）。以下分析针对一个框架进行，其中包含 $n = 4$, $\lambda = 10^{-6}$，并在 KSE $L = 22$ 数据集上训练；该数据集具有一个 64维环境空间，其内含一个具有 $d_m = 8$ 的非线性不变流形。

我们首先定义模型中从输入到输出的若干关键权重 W_j 和表示 z_j，其中 j 是网络中位置的占位符。从编码器开始，我们将编码器非线性部分直接输出的非线性激活表示 $\mathcal{E}_N$ 定义为 $z_{\mathcal{E}N} = \mathcal{E}_N(u)$。随后，该表示通过线性层 $W_\mathcal{E}$ 映射到 $\mathbb{R}^{d_z}$，得到表示 $z_\mathcal{E} = \mathcal{E}(u) = W_\mathcal{E}\mathcal{E}_N(u)$。接着，该表示依次通过 n 个方形线性层： $W_1, ..., W_n$。每经过一层后输出的表示分别为 $z_1, ..., z_n$。请注意，按照前几节的术语， z_n 等于 z。最后，在进入非线性解码器之前， $\mathcal{D}_N$, z_n 通过 $W_\mathcal{D}$ 映射到适当的大小： $\mathcal{D}(z_n) = \mathcal{D}_N(W_\mathcal{D} z_n)$。总而言之，数据快照经过完整编码和解码后为 $\tilde{u} = \mathcal{D}(\mathcal{W}(\mathcal{E}(u))) = \mathcal{D}_N(W_\mathcal{D} W_n ... W_1 W_\mathcal{E}\mathcal{E}_N(u))$。

我们首先通过计算数据表示的协方差z_j，来追踪训练过程中网络各个中间层和不同训练轮次下潜在表示的秩随时空的变化，如图15所示。沿着模型在空间（层）中的方向进行考察时，我们发现非线性编码器会产生满秩表示，并不直接负责将数据转换为低秩形式，如图15a所示。然而，我们观察到，当数据从非线性编码器出发，经过非方形线性映射到达 W_1时，学习到的表示

呈现弱低秩特性，如图15b所示。当数据依次经过各个方形线性块 W_1，.....， W_n时，我们观察到，表示中不必要的奇异值/方向会进一步衰减（等价地，最关键的表示方向会得到增强），使潜在表示逐渐趋向真正的最小秩表示。

随着我们沿时间维度（即训练轮次）考察模型，我们发现，为 $z_{\mathcal{E}N}$ 学习到的表示的秩停滞不变。相比之下，对于每个串联线性层，我们观察到其表示的秩起初几乎是满秩的，但随后逐渐共同衰减为低秩表示。我们注意到，在训练早期，表示会“过度修正”，变成秩过低、无法准确捕捉数据的表示；但随着训练的推进，网络会自动解决这一问题。

为了进一步了解其中发生的情况，我们现在对模型进行类似的时空分析，以追踪每一层权重的梯度更新的秩， $\mathcal{J}_j = \nabla_{W_j}\mathcal{L}$，如图16所示。这里我们沿用上文所述的相同层索引约定。从图16 中可以看出，在训练早期，串联线性层的更新梯度会调整潜在表示中的所有方向。随着训练推进，更新梯度的奇异值在不必要的方向上开始衰减，使潜在空间转向低秩表示。达到这一状态后，梯度更新基本上只会在重构所需的重要方向上进行更新。通过对梯度更新的分析，我们可以得出结论：该框架会共同调整所有线性层。

由于串联的线性层可以通过简单地计算其乘积合并为一个线性层，我们还研究了有效层权重矩阵本身的秩，$W_{j,\ \mathrm{eff}}$ (例如 $W_{2,\ \mathrm{eff}} = W_2W_1W_{\mathcal{E}}$)，其随空间和时间的变化如图17所示。我们在图17 中展示，随着线性层在网络中逐渐加深，这些层的有效秩趋近于 d_m。这与图15中的观察结果一致。由此我们得出结论：串联线性层协同形成一个有效秩为 d_m 的权重矩阵，将数据投影到维数为 d_m的空间中。需要指出的是，尽管网络自动学习到了线性分离的 d_m 表示，原始数据集的流形本质上是非线性的，并且非线性地嵌入在环境空间中——这一特征由非线性编码和解码块捕获。最后，我们评论

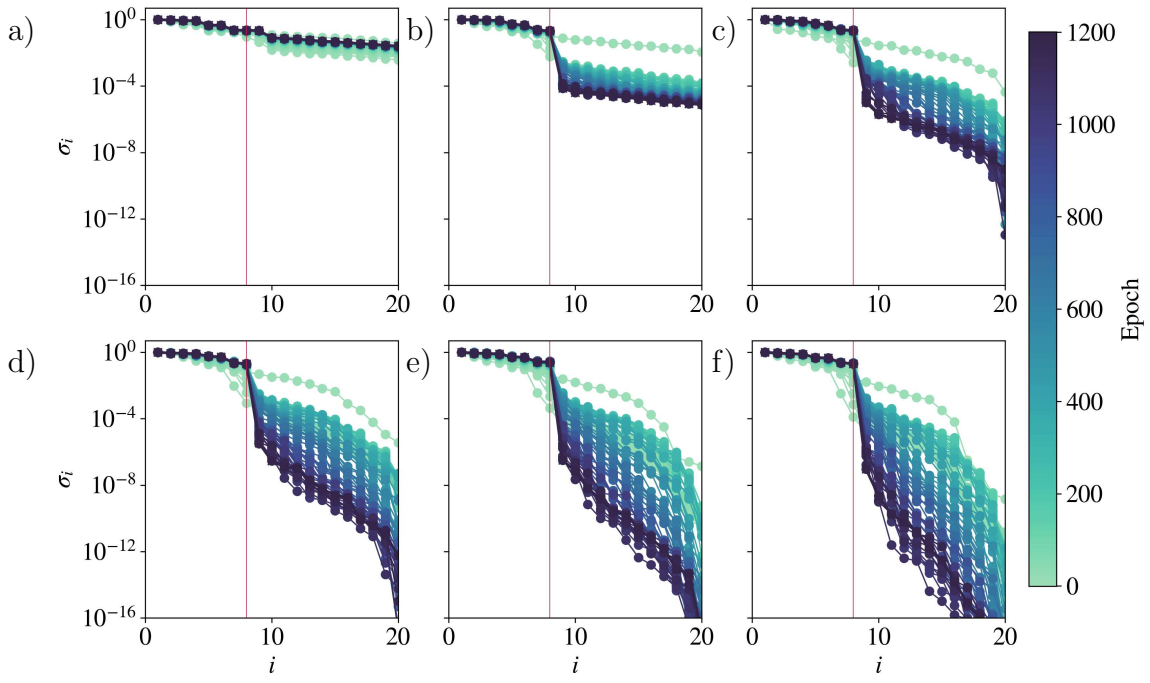

图 15：在 KSE $L = 22$ 数据集上训练的 数据表示的协方差奇异谱的“时空”跟踪：a) $z_{\mathcal{E}N}$ b) $z_{\mathcal{E}}$ c) z_1，d) z_2，e) z_3，以及 f) z_4 （即 z）随训练轮次的变化。请注意，为清晰起见，a) 和 b) 的谱已截断。数据集的 d_m 由一条竖直红线表示。

当在线性块之间实施权重共享时 W_j (即 W_j 相等)，我们会失去正则化，因为权重共享减少了线性层的有效数量。

本节最后比较我们提出的框架 IRMAE-WD（采用隐式正则化和权重衰减）与仅采用隐式正则化的 IRMAE。我们在图 18 中展示了两者潜在表示的数据协方差的学习动态。图 18a 展示了没有权重衰减时的动态：我们观察到，尾部奇异值在前 100 个训练轮次中先向上漂移，随后衰减，并在训练继续进行时再次增长。如图 18b所示，加入权重衰减后，尾部奇异值呈单调衰减。这些观察结果与附录 C中的线性 IRMAE-WD 分析一致；该分析预测，在没有权重衰减时，特征值为零的方向会发生漂移，训练动力学将在这些方向上

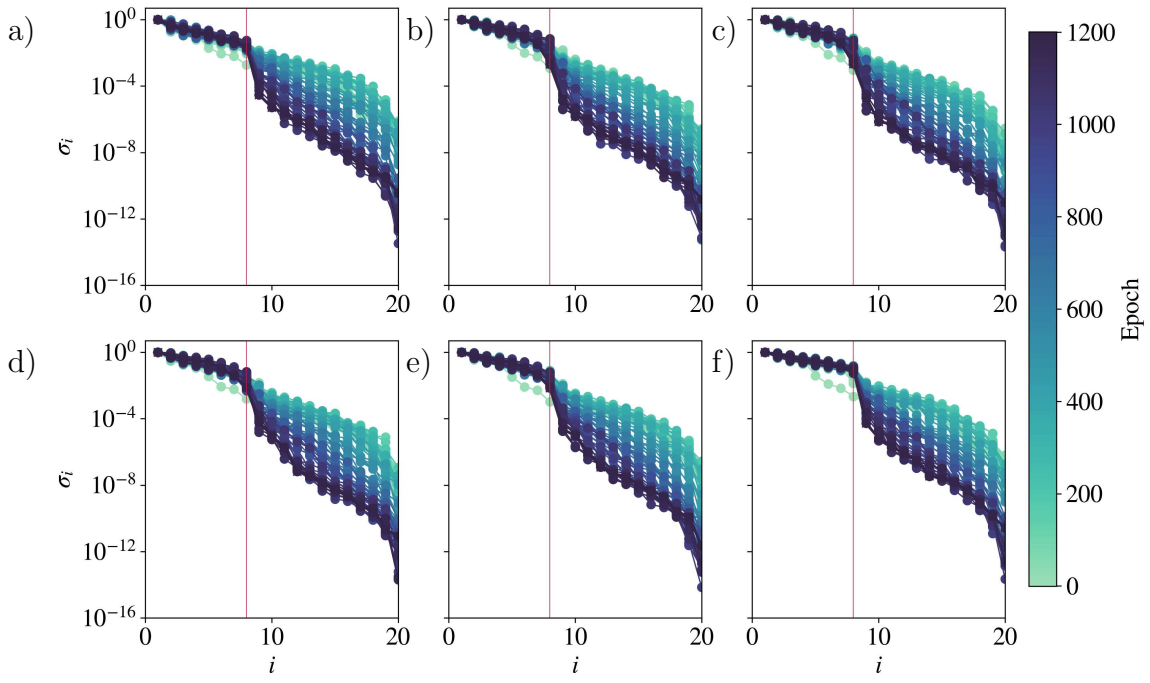

图16：对于在 KSE $L = 22$ 数据集上训练的模型，更新梯度 $\mathcal{J}_j$ 奇异谱的“时空”追踪，随训练轮次变化，分别对应 a) $\mathcal{J}_\mathcal{E}$ b) $\mathcal{J}_1$、c) $\mathcal{J}_2$、d) $\mathcal{J}_3$、e) $\mathcal{J}_4$ 和 f) $\mathcal{J}_\mathcal{D}$。请注意，为了清晰起见，a) 和 f) 的谱已截断。数据集的 d_m 用一条红色竖线表示。

漂移。加入权重衰减会使这些特征值变为负值，从而有助于收敛。

IV. 结论

在本文中，我们基于 Jing 等人的观察结果，[32] 并提出一种自编码器框架，称为 IRMAE-WD。该框架将隐式正则化与内部线性层和权重衰减相结合，以自动估计数据所在流形的内在维数 d_m。该框架能够同时学习有序且正交的流形坐标表示，以及环境空间与流形空间之间的映射函数，从而实现样本外投影。与其他自编码器方法不同，我们无需进行参数模型扫描，也不依赖于

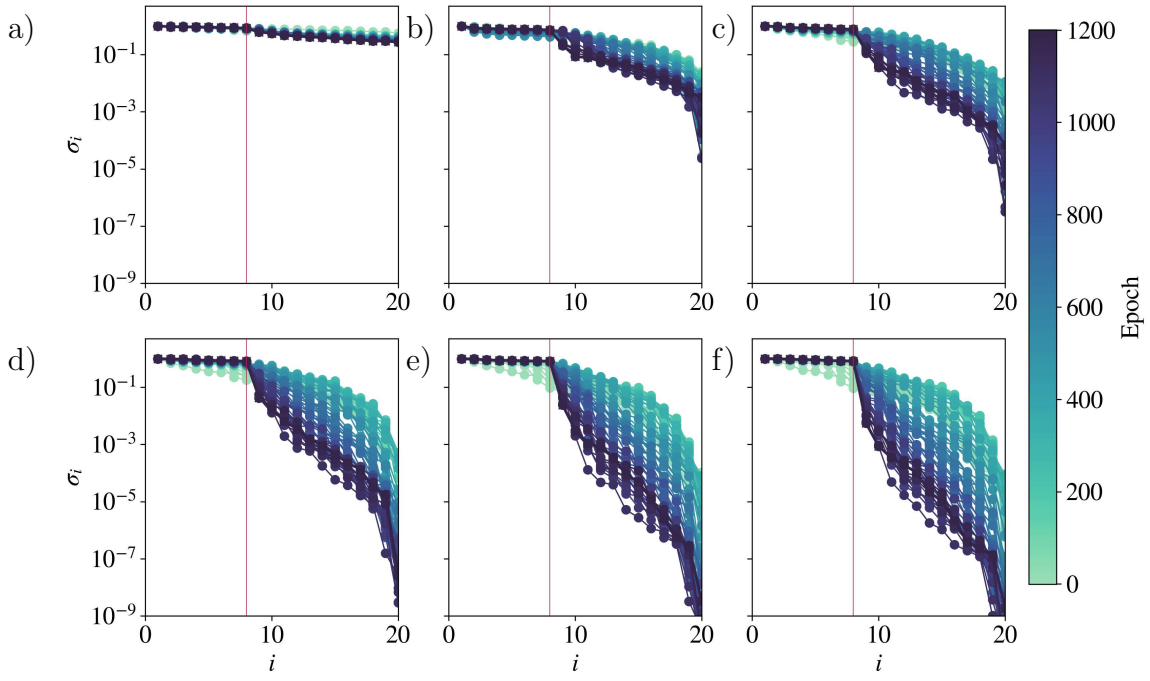

图17：对在 KSE 数据集上训练的模型，其有效线性层$W_{j,\ \text{eff}}$奇异谱的“时空”追踪。有效权重矩阵的奇异谱：a) $W_{\mathcal{E},\ \text{eff}}$ b) $W_{1,\ \text{eff}}$，c) $W_{2,\ \text{eff}}$，d) $W_{3,\ \text{eff}}$，e) $W_{4,\ \text{eff}}$，以及 f) $W_{\mathcal{D},\ \text{eff}}$，随训练轮次的变化。注意，为了清晰起见，a) 和 f) 的谱经过了截断。数据集的 d_m 以一条红色竖线表示。

辅助算法只要求自编码器的瓶颈维度 d_z 满足 $d_z > d_m$。

我们通过估计一系列具有线性和非线性流形的有限维系统及（离散化的）无限维系统的流形维数，展示了我们的框架。我们证明，对于具有非线性嵌入流形的系统，该框架优于若干最先进的估计器；即使流形维数相对较大，它仍能保持准确，$d_m \approx 40$。然而，与许多工业应用的需求相比，我们测试系统的环境维度仍然较小，例如湍流，其中环境维度 d_u（傅里叶模态数或网格点数）很容易超过 10^6，而且人们推测 d_m 会随雷诺数（流动强度）大幅增加。未来，我们将致力于高效地把 IRMAE-WD 扩展到这些高维系统。

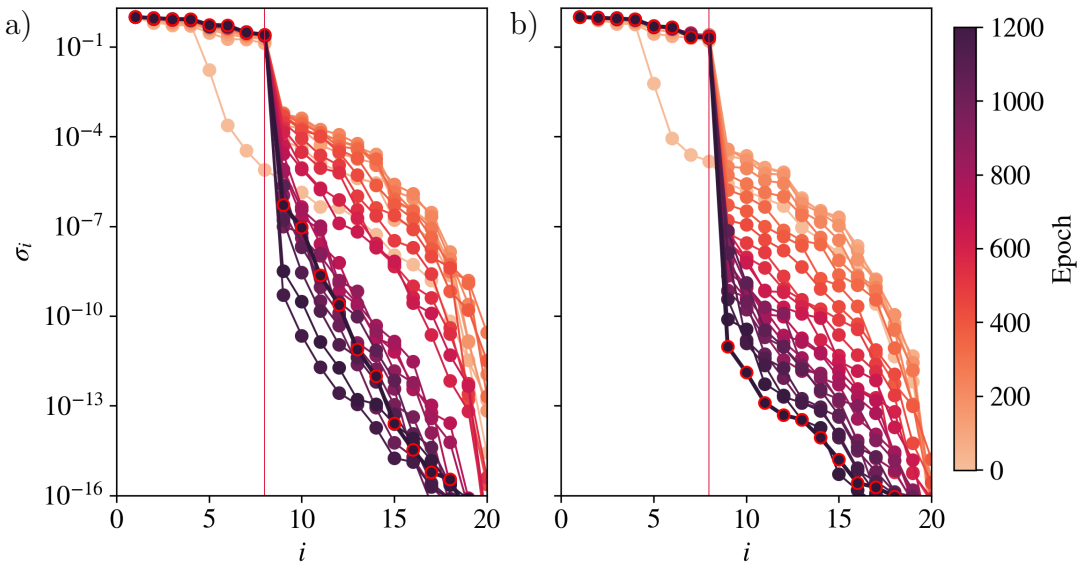

图18：对在 KSE $L = 22$ 数据集上训练的表示数据z的协方差奇异谱进行“时空”追踪：a) 仅采用隐式正则化的自编码器（IRMAE）；b) 采用隐式正则化和权重衰减的自编码器（IRMAE-WD）。数据集的 d_m 以一条红色竖线表示，最终学习到的潜在谱则以红色标记勾勒。

我们证明，我们的框架可以自然地扩展到使用 Kuramoto-Sivashinsky 方程和 lambda-omega 反应-扩散系统进行状态空间建模与预测的应用中。通过使用神经常微分方程，我们学习了这些数据集在流形表示中的动力学，并表明可以在任意所需时刻准确恢复环境空间表示。

我们对训练过程在“空间”（层）和“时间”（轮次）中的分析表明，低秩学习会同时出现在所有线性层中。我们强调，非线性编码器并不直接负责学习低秩表示；相反，每个串联线性层协同作用，通过叠加潜在空间中近似低秩的特征，有效增强相关的流形方向，并相应削弱冗余模态。对采用 IRMAE- WD 架构的线性自编码器进行分析，展示了线性层在加速编码器、解码器和内部层集体收敛中的作用，也说明了权重衰减如何打破退化，避免这些退化在缺少权重衰减时限制收敛。从理论角度来看，尽管附录C 中给出的线性自编码器分析提供了一些见解，但其局限性很大，而且

还需要开展更深入、更复杂的研究，以便更好地理解该方法，即使在线性情形下也是如此，更不用说完全非线性的情形了。

最后，我们证明，我们的框架对 L_2 正则化（权重衰减）参数 λ 和线性层数量 n的选择具有很强的鲁棒性。我们展示了，在很大范围的正则化参数下，我们都能准确估计 d_m ，且不会牺牲准确度（均方误差）。我们还表明， λ 有助于减少学习得到的潜在空间中多余奇异方向的影响。

虽然本研究的动机来自复杂的确定性动力系统，但我们也认识到，许多值得关注的实际系统具有随机性或噪声，数据可能仅近似位于有限维流形上，而非精确位于其上。未来，我们计划将 IRMAE-WD 稳健地扩展到这些系统。

致谢

本研究得到美国海军研究办公室资助，项目编号为 N00014-18-1-2865（范内瓦·布什教师奖学金）。我们还要感谢威斯康星大学麦迪逊分校工程学院工程研究学者研究生项目（GERS），并鸣谢高级机会奖学金（AOF）和 PPG 奖学金提供的资助。

数据可用性声明

支持本研究发现的代码和示例数据可在以下网址公开获取：
https://github.com/mdgrahamwisc/IRMAE_WD

附录 A：模型架构与参数

附录 B：在 MNIST 手写数字数据集上的应用

这里我们将 IRMAE-WD 应用于 MNIST 数据集，并与 Jing 等人进行比较。[32]我们采用了他们使用的相同卷积自编码器架构参数，具体参数和架构如下： 4×4 卷积核尺寸，步幅为 2，填充为 1，学习

表II：此处列出了本文研究中使用的架构和参数。为简洁起见，每种架构的解码器 $\mathcal{D}$ 只是编码器 $\mathcal{E}$的镜像，激活函数依次为 ReLU/ReLU/lin。每个网络在编码器和解码器之间有 n 个形状为$d_z \times d_z$ 的串联线性层。学习率设为 10^{-3}，小批量大小为 128。

数据集	$\mathcal{E}$	激活	d_z	n	λ
五维噪声	20/128/64/20	ReLU/ReLU/lin	20	4	10^{-2}
Arch. Lorenz	4/128/64/4	ReLU/ReLU/线性	4	4	10^{-6}
二维环面	3/256/128/10	ReLU/ReLU/线性	10	4	10^{-2}
KSE $L = 22$	64/512/256/20	ReLU/ReLU/lin	20	4	10^{-6}
KSE $L = 44$	64/512/256/30	ReLU/ReLU/lin	30	4	10^{-6}
KSE $L = 66$	64/512/256/50	ReLU/ReLU/lin	50	4	10^{-6}
KSE $L = 88$	20/512/256/80	ReLU/ReLU/lin	80	4	10^{-6}

率为 10^{-3}，以及 $\lambda = 10^{-6}$。这里的 Conv、ConvT 和 FC 分别对应卷积层、转置卷积层和全连接（未激活）层。

在图 19中，我们展示了 IRMAE-WD 同时利用隐式正则化和权重衰减，学到了 MNIST 手写数字数据集的 $d_m = 9$ 表示，而 Jing 等人[32], 仅利用隐式正则化，学到了 $d_m = 10$ 表示。我们还强调，与 Jing 等人的模型相比，我们的模型的尾部奇异值急剧衰减，低了几个数量级。最后，我们注意到，Jing 等人的[32] 模型的潜在空间呈现出更宽的尾部，尤其是在显著奇异值附近。我们发现，尽管我们的模型在对 MNIST 数据建模时少用了一个自由度，但其均方误差与 Jing 等人的模型相当 [32]，前提是使用他们的参数进行训练（ [32] 对比 $1.0 \cdot 10^{-2}$）。 $9.5 \cdot 10^{-3}$

表III：MNIST 手写数字数据集的卷积自编码器及架构

编码器	解码器
$x \in \mathbb{R}^{32\times32\times1}$	$z \in \mathbb{R}^{128}$
$\rightarrow$卷积$_{32}$ $\rightarrow$ ReLU	$\rightarrow$ FC$_{4096}$
$\rightarrow$ 卷积$_{64}$ $\rightarrow$ReLU	$\rightarrow$ 重塑$_{8\times8\times64}$
$\rightarrow$ 卷积 $_{128}$ $\rightarrow$ ReLU	$\rightarrow$转置卷积$_{64}$ $\rightarrow$ReLU
$\rightarrow$ 卷积 $_{256}$ $\rightarrow$ ReLU	$\rightarrow$转置卷积$_{32}$ $\rightarrow$ReLU
$\rightarrow$ 展平 $_{1024}$	$\rightarrow$卷积转置$_1$ $\rightarrow$双曲正切
$\rightarrow$LC$_{128}$ $\rightarrow$ $z \in \mathbb{R}^{128}$	$\rightarrow$ $x \in \mathbb{R}^{32\times32\times1}$

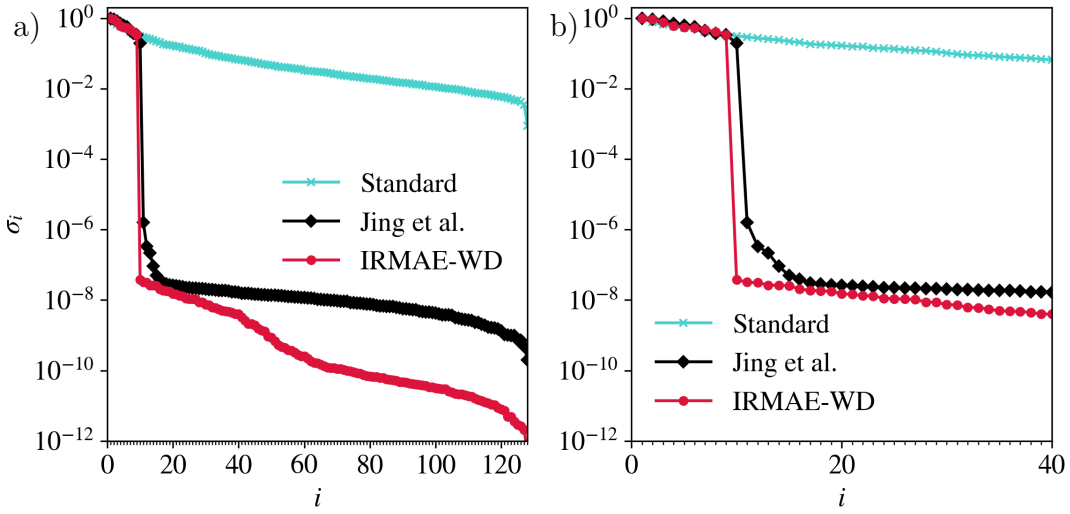

图19：在 MNIST 手写数字数据集上训练的模型所得的奇异值谱：a) 完整谱，b) 放大后的谱。

附录 C：带内部线性层和权重衰减的线性自编码器分析

1. 公式化

为了深入了解带有额外线性层和权重衰减的自编码器的性能，我们在此分析一个理想化情形下的梯度下降：具有一个或多个内部线性层的线性自编码器。我们首先从只有一个内部线性层的形式化开始。输入记为 $u \in \mathbb{R}^{d_u}$，编码器记为 $E \in \mathbb{R}^{d_z \times d_u}$，解码器记为 $D \in \mathbb{R}^{d_u \times d_z}$，内部线性层记为 $W \in \mathbb{R}^{d_z \times d_z}$，输出记为 $\tilde{u} = DWEu \in \mathbb{R}^{d_u}$。我们将线性层之前的潜变量定义为 $h = Eu \in \mathbb{R}^{d_z}$，之后的潜变量定义为$w = Wh = WEu \in \mathbb{R}^{d_z}$。对于传统自编码器，$W = I^{d_z d_z}$，其中记号$I^{mm}$ 表示 $m \times m$ 单位矩阵。我们将考虑简单的损失函数

$$\mathcal{L} = \langle ||\tilde{u} - u||_2^2 \rangle + \lambda_E (||E||_F^2 + ||D||_F^2) + \lambda_W ||W||_F^2,$$

其中，$\langle \cdot \rangle$表示集成平均（期望值）。首先将考察损失最小化问题的收敛平衡态解，然后讨论该解向最小值的收敛

使解达到最小值。

我们尤其关注这样一种情况：数据位于 $\mathbb{R}^{d_u}$ 的一个 r 维子空间上，或者等价地，秩为 $\langle uu^T\rangle = r$；并假设隐藏层的维数 m 选取得满足 $m > r$。

我们可以将损失写为

$$\mathcal{L} = \langle u^T(E^T W^T D^T DWE - (DWE + E^T W^T D^T))u\rangle + \langle u^T u\rangle + \lambda_E(\mathrm{tr}EE^T + \mathrm{tr}DD^T) + \lambda_W \mathrm{tr}WW^T.$$

用指标记号可以写为

$$\begin{aligned}\mathcal{L} &= \langle u_k E^T_{kl} W^T_{lm} D^T_{mn} D_{no} W_{op} E_{pq} u_q\rangle \\ &\quad - \langle u_k (D_{kl} W_{lm} E_{mn} + E^T_{kl} W^T_{lm} D^T_{mn}) u_n\rangle \\ &\quad + \langle u_k u_k\rangle + \lambda_E(E_{kl}E_{kl} + D_{kl}D_{kl}) + \lambda_W W_{kl} W_{kl}.\end{aligned}$$

求偏导数可得

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial E_{ij}} &= \langle 2W^T_{im} D^T_{mn}(DWE - I)_{no} u_o u_j\rangle + 2\lambda_E E_{ij} = \langle 2W^T_{im} D^T_{mn}(\tilde{u}_n - u_n) u_j\rangle + 2\lambda_E E_{ij}, \\ \frac{\partial \mathcal{L}}{\partial W_{ij}} &= \langle 2D^T_{ik}(DWE - I)_{kl} u_l E_{jm} u_m\rangle + 2\lambda W_{ij} = \langle 2D^T_{ik}(\tilde{u}_k - u_k) h_j\rangle + 2\lambda_E W_{ij}, \\ \frac{\partial \mathcal{L}}{\partial D_{ij}} &= \langle 2(DWE - I)_{ik} u_k (WE)_{jl} u_l\rangle + 2\lambda_E D_{ij} = 2(\tilde{u}_i - u_i) w_j + 2\lambda_E D_{ij}.\end{aligned}$$

我们考虑一个高度理想化的数据集，其中 $\langle uu^T\rangle = \sigma^2 I^{rd_u d_u}$，其中 I^{rpq} 是一个 $p \times q$ 矩阵（具有 p、$q > r$），其前 r 个对角元素为 1，其余元素均为 0。现在

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial E_{ij}} &= 2\sigma^2 W^T_{im} D^T_{mn}(DWE - I)_{no} I^{rd_u d_u}_{oj} + 2\lambda_E E_{ij}, \\ \frac{\partial \mathcal{L}}{\partial W_{ij}} &= 2\sigma^2 D^T_{ik}(DWE - I)_{kl} E_{jm} I^{rd_u d_u}_{lm} + 2\lambda_W W_{ij}, \\ \frac{\partial \mathcal{L}}{\partial D_{ij}} &= 2\sigma^2 (DWE - I)_{ik}(WE)_{jl} I^{rd_u d_u}_{kl} + 2\lambda_E D_{ij}.\end{aligned}$$

用矩阵-向量表示法可写为

$$\frac{\partial \mathcal{L}}{\partial E} = 2\sigma^2 W^T D^T (DWE - I) I^{rd_u d_u} + 2\lambda_E E,$$
$$\frac{\partial \mathcal{L}}{\partial W} = 2\sigma^2 D^T (DWE - I) I^{rd_u d_u} E^T + 2\lambda_W W,$$
$$\frac{\partial \mathcal{L}}{\partial D} = 2\sigma^2 (DWE - I) I^{rd_u d_u} (WE)^T + 2\lambda_E D.$$

2. 平衡解

收敛时，这些导数为零。暂且令 $\lambda_E = 0$。我们先考虑不存在内部线性层时的解，即 $W = I^{d_z d_z}$ 时的情况。此时，当满足以下条件时， $\frac{\partial \mathcal{L}}{\partial E}$ 和 $\frac{\partial \mathcal{L}}{\partial D}$ 将为零

$$(DE - I) I^{rd_u d_u} = DEI^{rd_u d_u} - I^{rd_u d_u} = 0.$$

这有一个“满秩”解 $E = D^T = I^{d_z d_u}$，满足 $DE - I = 0$，还有一个“秩 r”解 $E = D^T = I^{rd_z d_u}$。它不满足 $DE - I = 0$，但满足$DEI^{rd_u d_u} - I^{rd_u d_u} = 0$。若加入一个非平凡的线性层 W，则有

$$(DWE - I) I^{rd_u d_u} = DWEI^{rd_u d_u} - I^{rd_u d_u} = 0.$$

显然，秩为 r 的解 $E = D^T = I^{rd_u d_z}$，以及秩为 r 的选取 $W = I^{rd_z d_z}$，仍然是解；同样，满足 $E = D^T = I^{d_z d_u}$ 、且 $W = I^{d_z d_z}$ 的满秩解也是如此。

存在权重衰减时，情况会更复杂，我们将只考虑 $\lambda_E = \lambda_W = \lambda$ 的情形。定义一个新参数 $\zeta = \lambda/\sigma^2$，并令该参数较小，可得如下形式的摄动解

$$E = I^{rd_z d_u}(1 + \alpha\zeta + O(\zeta^2)), W = I^{rd_z d_z}(1 + \beta\zeta + O(\zeta^2)), D = I^{rd_u d_z}(1 + \gamma\zeta + O(\zeta^2)) \quad \text{(C1)}$$

可以找到。 代入平衡条件 $\frac{\partial \mathcal{L}}{\partial E} = 0, \frac{\partial \mathcal{L}}{\partial W} = 0, \frac{\partial \mathcal{L}}{\partial D} = 0$ 并

忽略 $O(\zeta^2)$ 项后，每种情况下均得到

$$\alpha + \beta + \gamma + 1 = 0.$$

因此，带权重衰减的平衡问题存在一整族解。为便于后文引用，我们将这个解（相差 $O(\zeta)$）写作

$$\begin{aligned} E &= aI^{rd_z d_u}, \quad a = 1 + \alpha\zeta, \\ W &= bI^{rd_z d_z}, \quad b = 1 + \beta\zeta, \\ D &= cI^{rd_u d_z}, \quad c = 1 + \gamma\zeta. \end{aligned} \tag{C2}$$

3. 梯度下降的收敛

a. 动力学模型

现在我们转而讨论梯度下降如何收敛到平衡解的问题。在这里，我们考虑一个非常简单的常微分方程模型来描述这一过程，其中 t 是伪时间，表示梯度下降步数：

$$\begin{aligned} \frac{dE}{dt} &= -\frac{1}{2\sigma^2}\frac{\partial \mathcal{L}}{\partial E} = -W^T D^T (DWE - I) I^{rd_u d_u} - \zeta E, \\ \frac{dW}{dt} &= -\frac{1}{2\sigma^2}\frac{\partial \mathcal{L}}{\partial W} = -D^T (DWE - I) I^{rd_u d_u} E^T - \zeta W, \\ \frac{dD}{dt} &= -\frac{1}{2\sigma^2}\frac{\partial \mathcal{L}}{\partial D} = -(DWE - I) I^{rd_u d_u} (WE)^T - \zeta D. \end{aligned} \tag{C3}$$

这是一个高维且高度非线性的系统；为了取得进展，我们仅考虑收敛附近的动力学，并在收敛解周围对系统进行线性化（C1）。也就是说，我们令

$$\begin{aligned} E &= aI^{rd_z d_u} + \epsilon \hat{E}, \\ W &= bI^{rd_z d_z} + \epsilon \hat{W}, \\ D &= cI^{rd_u d_z} + \epsilon \hat{D}, \end{aligned} \tag{C4}$$

ˆˆˆ其中 E, W, D 是偏离收敛解的扰动。将这些表达式代入（C3），并忽略 O（ϵ^2）阶的项，得到

$$\begin{aligned}
\frac{d\hat{E}}{dt} &= -bc\left[abI^{rd_zd_u}\hat{D}I^{rd_zd_u} + acI^{rd_zd_z}\hat{W}I^{rd_zd_u} + bcI^{rd_ud_z}\hat{E}I^{rd_zd_z}\right] - \zeta\hat{E}, \\
\frac{d\hat{W}}{dt} &= -ab\left[abI^{rd_zd_u}\hat{D}I^{rd_zd_z} + acI^{rd_zd_z}\hat{W}I^{rd_zd_z} + bcI^{rd_zd_z}\hat{E}I^{rd_ud_z}\right] - \zeta\hat{W}, \\
\frac{d\hat{D}}{dt} &= -ac\left[abI^{rd_ud_z}\hat{D}I^{rd_zd_z} + acI^{rd_ud_z}\hat{W}I^{rd_zd_z} + bcI^{rd_ud_z}\hat{E}I^{rd_ud_z}\right] - \zeta\hat{D}.
\end{aligned} \tag{C5}$$

现在我们可以对这些解作出一些重要的概括性陈述。首先注意，方括号中的项总会得到只有左上角 $r \times r$ 块非零的矩阵。此外，对于任意非零的 ζ，该块之外的所有项都会趋于零。最后注意，(C5) 将具有如下形式的时变解：ˆ$E(t) = \mathcal{E}(t)I^{rd_zd_u}$、ˆ$W(t) = \mathcal{W}(t)I^{rd_zd_z}$、ˆ$D(t) = \mathcal{D}(t)I^{rd_ud_z}$，其中 $\mathcal{E}$、$\mathcal{D}$ 和 $\mathcal{W}$ 是时间的标量函数。这些扰动的演化方程为

$$\begin{aligned}
\frac{d\mathcal{E}}{dt} &= -bc\left[ab\mathcal{D} + ac\mathcal{W} + bc\mathcal{E}\right] - \zeta\mathcal{E}, \\
\frac{d\mathcal{W}}{dt} &= -ab\left[ab\mathcal{D} + ac\mathcal{W} + bc\mathcal{E}\right] - \zeta\mathcal{W}, \\
\frac{d\mathcal{D}}{dt} &= -ac\left[ab\mathcal{D} + ac\mathcal{W} + bc\mathcal{E}\right] - \zeta\mathcal{D}.
\end{aligned} \tag{C6}$$

下文中，我们将考虑这一不变子空间中的解，在该子空间内，可以对线性化动力学作出相当完整的刻画。

b. 线性层加速权重的集体收敛

没有权重衰减时，情况最简单：$\zeta = 0$。现在 $a = b = c = 1$ 且（C6）简化为

$$\begin{aligned}
\frac{d\mathcal{E}}{dt} &= -\left[\mathcal{D} + \mathcal{W} + \mathcal{E}\right], \\
\frac{d\mathcal{W}}{dt} &= -\left[\mathcal{D} + \mathcal{W} + \mathcal{E}\right], \\
\frac{d\mathcal{D}}{dt} &= -\left[\mathcal{D} + \mathcal{W} + \mathcal{E}\right].
\end{aligned} \tag{C7}$$

将这些方程相加可得:

$$\frac{d}{dt}(\mathcal{D}+\mathcal{W}+\mathcal{E})=-3(\mathcal{D}+\mathcal{W}+\mathcal{E}).$$

因此，“集体”权重（扰动） $\mathcal{C}=\mathcal{D}+\mathcal{W}+\mathcal{E}$ 按如下方式衰减： $e^{-\rho_1 t}$ ，其中 $\rho_1=3$。

我们还可以求出损失的演化方程。对于较小的 ϵ,

$$\mathcal{L}=r\sigma^2\epsilon^2(\mathcal{D}+\mathcal{W}+\mathcal{E})^2. \tag{C8}$$

也就是说，损失与集体权重 $\mathcal{C}$的平方成正比。由于 $\mathcal{C}(t)=\mathcal{C}(0)e^{-\rho_1 t}$ ，我们便得到

$$\mathcal{L}(t)=\mathcal{L}(0)e^{-2\rho_1 t}. \tag{C9}$$

更一般地，我们可以将C7写成矩阵向量形式

$$\frac{d}{dt}\begin{bmatrix}\mathcal{E}\\ \mathcal{W}\\ \mathcal{D}\end{bmatrix}=A\begin{bmatrix}\mathcal{E}\\ \mathcal{W}\\ \mathcal{D}\end{bmatrix},\qquad A=-\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}. \tag{C10}$$

其通解为

$$\begin{bmatrix}\mathcal{E}(t)\\ \mathcal{W}(t)\\ \mathcal{D}(t)\end{bmatrix}=C_1e^{-3t}v_1+C_2v_2+C_3v_3$$

with

$$v_1=\frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\end{bmatrix},\qquad v_2=\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\0\\1\end{bmatrix},\qquad v_3=\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\\0\end{bmatrix},$$

以及 $C_i=[\mathcal{E}(0)\ \mathcal{W}(0)\ \mathcal{D}(0)]^T v_i$。因此，虽然集体权重 $\mathcal{D}+\mathcal{W}+\mathcal{E}$按$e^{-\rho_1 t}$衰减，损失按 $e^{-2\rho_1 t}$衰减，但量 $\mathcal{D}-\mathcal{E}$和 $\mathcal{W}-\mathcal{E}$完全不会衰减，这是因为矩阵 G有两个零特征值。这一事实将限制梯度的性能。

在不存在权重衰减的情况下，生成低秩权重矩阵时的下降。（我们将在下文看到，权重衰减会打破动力学的简并性。）

现在，我们来讨论内部线性层的数量如何影响收敛。要考虑没有内部线性层的情况，只需将 W 以及 $\mathcal{W}$ 设为零——矩阵 W 固定为单位矩阵。此时，(C10) 化简为

$$\frac{d}{dt}\begin{bmatrix}\mathcal{E}\\ \mathcal{D}\end{bmatrix} = -\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\mathcal{E}\\ \mathcal{D}\end{bmatrix}. \tag{C11}$$

现在，集体权重变量 $\mathcal{D}+\mathcal{E}$ 以 $e^{-\rho_0 t}$衰减，其中 $\rho_0 = 2$；而在存在一个内部线性层时，它以 e^{-3t} 衰减——这一新增层加速了沿集体特征方向的收敛。

如果我们添加额外的线性层，使总数达到 n，即将 W 替换为乘积 $W_n W_{n-1} W_{n-2} \cdots W_1$，会怎样？不失一般性地，我们可以取这些矩阵中的每一个在收敛后的值（不存在权重衰减时）为 $I^{rd_z d_z}$。在考察线性化动力学时，我们使用如下结果

$$\begin{aligned} W_n W_{n-1} W_{n-2} \cdots W_1 &= (I^{rd_z d_z} + \epsilon \hat{W}_n)(I^{rd_z d_z} + \epsilon \hat{W}_{n-1})(I^{rd_z d_z} + \epsilon \hat{W}_{n-2}) \cdots (I^{rd_z d_z} + \epsilon \hat{W}_1) \\ &= I^{rd_z d_z} + \epsilon(\hat{W}_n + \hat{W}_{n-1} + \hat{W}_{n-2} + \cdots + \hat{W}_1) + O(\epsilon^2). \end{aligned}$$

取 $\hat{W}_i = \mathcal{W}_i I^{rd_z d_z}$ 并按照上述相同步骤，可得到线性化动力学的以下方程组：

$$\frac{d}{dt}\begin{bmatrix}\mathcal{E}\\ \mathcal{W}_n\\ \mathcal{W}_{n-1}\\ \vdots\\ \mathcal{W}_1\\ \mathcal{D}\end{bmatrix} = -\begin{bmatrix}1 & 1 & 1 & \cdots & 1 & 1\\ 1 & 1 & 1 & \cdots & 1 & 1\\ 1 & 1 & 1 & \cdots & 1 & 1\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 1 & 1 & 1 & \cdots & 1 & 1\end{bmatrix}\begin{bmatrix}\mathcal{E}\\ \mathcal{W}_n\\ \mathcal{W}_{n-1}\\ \vdots\\ \mathcal{W}_1\\ \mathcal{D}\end{bmatrix} \tag{C12}$$

通过添加这些　　结合这些方程，我们发现此情况下的集体权重为 $\mathcal{C} =$

$\mathcal{D} + \sum_{i=1}^{n} \mathcal{W}_i + \mathcal{E}$ 按 $e^{-\rho_n t}$衰减， n 层网络的衰减率为 ρ_n。

$$\rho_n = 2 + n. \tag{C13}$$

同样，增加内部层后，损失仍与 $\mathcal{C}^2$ 成正比，因此以 $2\rho_n$ 的速率衰减。与没有内部层的情况相比， $\mathcal{C}$ 和 $\mathcal{L}$ 的衰减率为

$$\frac{\rho_n}{\rho_0} = 1 + \frac{n}{2}. \tag{C14}$$

集体权重变量 $\mathcal{C}$ （以及损失）收敛速度提高的根本原因，在于自编码器损失的基本结构——对于每一层，都会出现组合 $DWE - I$，因此所有层的梯度都具有共同的结构，其中包含集体权重 $\mathcal{C}$。内部线性层越多，这个集体权重收敛得就越快。

为通过一个示例说明这一结果，我们考虑了一个均值为零、协方差为 $\sigma^2 I^{rd_u d_u}$ 的数据集，其中 $r = 5$ 且 $d_u = 100$，并使用维数为 $d_z = 20$ 的内部层。我们对权重矩阵的所有对角元素施加小幅零均值噪声，使其偏离平衡态值（非对角元素保持为零），并从这一初始条件开始进行梯度下降。尽管这种扰动相当特殊，但它比上文规定的扰动更一般（在后者中，每个矩阵的所有对角元素都会受到相同幅度的扰动）。损失相对于初始值归一化后的演化如图 20 所示；衰减率与公式 C14 的标度关系完全一致。

现在有 $n + 1$ 个零特征值，表示梯度下降不起作用的方向： $\mathcal{D} - \mathcal{E}$ 和 $\mathcal{W}_i - \mathcal{E}$、$i = 1$、$\ldots n$。因此，虽然添加内部线性层加快了集体权重的收敛，但这些退化方向仍然存在，从而限制了梯度下降过程的整体性能。

c. 权重衰减打破简并性并带来渐近稳定性

如上文给出的平衡解结果所示，加入权重衰减会使分析复杂得多。因此，我们将仅限于

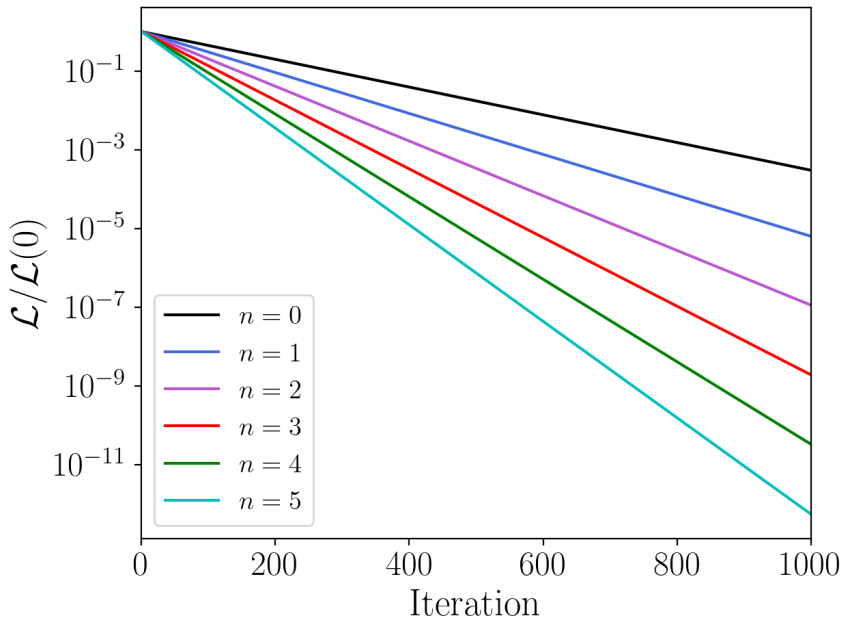

图2 0：对角扰动下线性网络梯度下降过程中相对损失（均方误差）的演化。

对以下情形的动力学进行微扰处理： $n = 1$ 当 ζ 较小时。将 a、b 和 c 的表达式代入（C6），并按 ζ 的同次幂合并，得到方程

$$\frac{d}{dt}\begin{bmatrix}\mathcal{E}\\ \mathcal{W}\\ \mathcal{D}\end{bmatrix} = (A+\zeta B)\begin{bmatrix}\mathcal{E}\\ \mathcal{W}\\ \mathcal{D}\end{bmatrix}, \tag{C15}$$

其中 A 与（C6）中相同，且

$$B = -\begin{bmatrix}2(\beta+\gamma)+1 & \gamma-1 & \beta-1\\ \gamma-1 & 2(\alpha+\gamma)+1 & \alpha-1\\ \beta-1 & \alpha-1 & 2(\alpha+\beta)+1\end{bmatrix}.$$

寻找形式为 $ve^{\xi t}$ 的解会导出特征值问题。

$$(A+\zeta B)v = \xi v.$$

对于较小的 ζ [43]，可以通过微扰方法求解。将特征向量 $v = v^{(0)} + v^{(1)} + O(\zeta^2)$ 和特征值 $\xi = \xi^{(0)} + \zeta\xi^{(1)} + O(\zeta^2)$ 展开后，可得到主阶问题

$$Ax^{(0)} = \xi^{(0)}x^{(0)} \tag{C16}$$

以及 $O(\zeta)$ 问题

$$(A - \xi^{(0)}I)v^{(1)} = (B - \xi^{(1)}I)v^{(0)}. \tag{C17}$$

主阶问题 (C16) 正是上文所述的无权重衰减情形，其特征值为 $\xi_1^{(0)} = -\rho_1 = -3$，$\xi_2^{(0)} = 0$，$\xi_3^{(0)} = 0$，特征向量为

$$v_1^{(0)} = \frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\end{bmatrix}, \qquad v_2^{(0)} = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\0\\1\end{bmatrix}, \qquad v_3^{(0)} = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\\0\end{bmatrix},$$

$O(\zeta)$ 问题是一个左端奇异的非齐次线性系统。给定一对特征值和特征向量 $\xi_i^{(0)}, v_i^{(0)}$，只有当右端项属于 $(A - \xi_i^{(0)}I)$ 的值域时，该系统才有解；等价地说，右端项必须与 $(A - \xi_i^{(0)}I)$ 的零空间正交 ^{T}。由于 $(A - \xi_i^{(0)}I)$ 是对称的，对于特征值 $\xi^{(0)} = \xi_i^{(0)}$，$(A - \xi_i^{(0)}I)$ 的零空间由 $v_i^{(0)}$ 张成，且当 $\left(v_i^{(0)}\right)^T (B - \xi^{(1)}I)v_i^{(0)} = 0$ 时方程有解。通过求解以下方程，可以确定 i 个特征值中的第 i 个特征值的 $O(\zeta)$ 修正量 $\xi_i^{(1)}$：

$$\xi_i^{(1)} = \frac{\left(v_i^{(0)}\right)^T B v_i^{(0)}}{\left(v_i^{(0)}\right)^T v_i^{(0)}}. \tag{C18}$$

计算后得到 $\xi_1^{(1)} = 3, \xi_2^{(1)} = \xi_3^{(1)} = -1$ （对于满足$\alpha + \beta + \gamma + 1 = 0$的任意 α, β, γ 选择均成立），因此误差为 $O(\zeta^2)$，我们有

$$\xi_1 = -3 + 3\zeta, \quad \xi_2 = \xi_3 = -\zeta. \tag{C19}$$

加入权重衰减会对集体收敛速率 $-\xi_1$ 产生非常轻微的不利影响，但更重要的是，它会将零处的特征值转换为负特征值，从而使系统沿所有方向衰减并趋近平衡态——平衡解变为

渐近稳定。

---

[1] 埃伯哈德·霍普夫。展示湍流特征的数学示例。《纯粹与应用数学通讯》，1(4)：303–322，1948年12月。doi:10.1002/cpa.3160010401。[2] 罗杰·特马姆。《力学与物理学中的无限维动力系统》。施普林格，纽约州纽约，1977年。[3] 谢尔盖·泽利克。《吸引子：过去与现在》。arXiv，2022年。doi:10.48550/arxiv.2208.12101。[4] 约翰·M·李。《光滑流形导论》，施普林格纽约丛书第218卷。施普林格纽约，2012年。国际标准书号 978-1-4419-9981-8。doi:10.1007/978-1-4419-9982-5。

[5] 丹尼尔·弗洛里安和Michael D. Graham。内在动力学的数据驱动发现。《自然·机器智能》，4(12):1113–1120，2022年。doi:10.1038/s42256-022-00575-4。网址 https://doi.org/10.1038/s42256-022-00575-4。

[6]Hong-liu Yang、Kazumasa A. Takeuchi、Francesco Ginelli、Hugues Chat´e和G¨unter Radons。空间扩展耗散系统的双曲性与有效维数。《物理评论快报》，102:074102，2009年2月。doi:10.1103/PhysRevLett.102.074102。网址 https://link.aps.org/doi/10.1103/PhysRevLett.102.074102。

[7] Hong-liu Yang和G¨unter Radons。通过李雅普诺夫投影法探测惯性流形的几何。《物理评论快报》，108:154101，2012年4月。doi:10.1103/PhysRevLett.108.154101。网址 https://link.aps.org/doi/10.1103/PhysRevLett.108.154101。

[8]丁 X、H. Chat´e、P. Cvitanovi´c、埃夫斯塔西奥斯·西米诺斯和 K. A. Takeuchi。从不稳定周期轨道估计惯性流形的维数。《物理评论快报》，117：024101，2016年7月。doi:10.1103/PhysRevLett.117.024101。网址 https://link.aps.org/doi/10.1103/PhysRevLett.117.024101。

[9] I. T. 乔利夫。 主成分分析。施普林格出版社，1986年。[10]邹、特雷弗·哈斯蒂和罗伯特·蒂布希拉尼。稀疏主成分分析。 计算与图形统计学杂志，15(2):262 – 286，2004年。网址=2-s2.0-29544445896&partnerID=40&m

d5=a6093fb8d8

[11]Christopher Bishop。《贝叶斯主成分分析》。载于 M. Kearns、S. Solla 和 D. Cohn（编），《神经信息处理系统进展》第 11 卷。麻省理工学院出版社，1998 年。网址

https://proceedings.neurips.cc/paper/1998/file/c88d8d0a6097754525e02c2246d8d27f-Paper.pdf.
[12]Juha Karhunen 和 Jyrki Joutsensalo。使用非线性主成分分析类型学习进行信号表示与分离。《神经网络》，7(1):113–127，1994。国际标准书号 0893-6080。doi:https://doi.org/10.1016/0893-6080(94)90060-4。URL https://www.sciencedirect.com/science/article/pii/0893608094900604。[13] Alec J. Linot 和 Michael D. Graham。通过深度学习发现并预测惯性流形上的动力学。《物理评论E》，101:062209，2020年6月。doi:10.1103/PhysRevE.101.062209。URL https://link.aps.org/doi/10.1103/PhysRevE.101.062209。[14] Alec J. Linot 和 Michael D. Graham。利用神经常微分方程对时空混沌进行数据驱动的降阶建模。《混沌：非线性科学跨学科期刊》，32(7)，2022年7月。国际标准书号 1054-1500。doi:10.1063/5.0069536。URL https://doi.org/10.1063/5.0069536。073110。[15] Pantelis R. Vlachas、Georgios Arampatzis、Caroline Uhler 和 Petros Koumoutsakos。通过学习有效动力学对复杂系统进行多尺度模拟。《自然·机器智能》，4(4):359–366，2022年。[16] Boyuan Chen、Kuang Huang、Sunand Raghupathi、Ishaan Chandratreya、Qiang Du 和 Hod Lipson。自动发现实验数据中隐藏的基本变量。《自然·计算科学》，2(7):433–442，2022年。[17]Carlos E. P´erez De Jes´us 和 Michael D. Graham。Kolmogorov流的数据驱动低维动力学模型。《物理评论：流体》，8(4):044402，2023年。doi: 10.1103/physrevfluids.8.044402。[18] Alec J. Linot 和 Michael D. Graham。湍流最小库埃特流的数据驱动低维模型的动力学。《流体力学杂志》，973:A42，2023年。国际标准书号 0022-1120。doi:10.1017/jfm.2023.720。[19]Matthew B. Kennel、Reggie Brown 和 Henry D. I. Abarbanel。利用几何构造确定相空间重构的嵌入维数。《物理评论A》，45:3403–3411，1992年3月。doi:10.1103/PhysRevA.45.3403。URL https://link.aps.org/doi/10.1103/PhysRevA.45.3403。[20] William Gilpin。从时间序列深度重构奇异吸引子。载于第34届神经信息处理系统国际会议论文集，NIPS'20，纽约州雷德

霍克，纽约州，美国，2020年。Curran Associates Inc. 国际标准书号 9781713829546。
[21] Zhe Wang 和 Claude Guet。基于含噪时间序列对神经动力系统进行自洽学习。IEEE计算智能新兴主题汇刊，6(5):1103–1112，2022年。doi:10.1109/TETCI.2022.3146332。

[22]Mikhail Belkin 和 Partha Niyogi。用于降维和数据表示的拉普拉斯特征映射。《神经计算》，15(6):1373–1396，2003年。doi：10.1162/089976603321780317。

[23]Laurens van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。机器学习研究期刊，9:2579–2605，2008年。URL：http://www.jmlr.org/papers/v9/vandermaaten08a.html。[24] Joshua B. Tenenbaum、Vin de Silva 和 John C. Langford。用于非线性降维的全局几何框架。《科学》，290(5500):2319–2323，2000年。doi:10.1126/science.290.5500.2319。URL：https://www.science.org/doi/abs/10.1126/science.290.5500.2319。[25] Sam T. Roweis 和 Lawrence K. Saul。通过局部线性嵌入进行非线性降维。《科学》，290(5500):2323–2326，2000年。doi:10.1126/science.290.5500.2323。URL：http://www.sciencemag.org/cgi/content/abstract/290/5500/2323。[26]Francesco Camastra 和 Antonino Staiano。内在维数估计：进展与开放问题。信息科学，328:26–41，2016年。ISSN 0020-0255。doi:https://doi.org/10.1016/j.ins.2015.08.029。URL：https://www.sciencedirect.com/science/article/pii/S0020025515006179。[27]Elizaveta Levina 和 Peter Bickel。内在维数的最大似然估计。载于 L. Saul、Y. Weiss 和 L. Bottou 编，神经信息处理系统进展，第17卷。麻省理工学院出版社，2004年。URL：https://proceedings.neurips.cc/paper/2004/file/74934548253bcab8490ebd74afed703

1-Paper.pd f.

[28]Anna V. Little、Yoon-Mo Jung 和 Mauro Maggioni。《数据集内在维数的多尺度估计》。第 FS-09-04 卷，第 26–33 页，2009 年。网址 https://www.scopus.com/inward/record.uri?eid=2-s2.0-77954233466&partnerID=40&m

d5=c0e2e6d98

[29] Suriya Gunasekar、Blake E Woodworth、Srinadh Bhojanapalli、Behnam Neyshabur 和 Nati Srebro。《矩阵分解中的隐式正则化》。载于 I. Guyon、U. V. Luxburg,

S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编），《神经信息处理系统进展》，第 30 卷。Curran Associates, Inc.，2017年。网址：https://proceedings.neurips.cc/paper/2017/file/58191d2a914c6dae66371c9dcdc91b41-Paper.pdf.[30] Sanjeev Arora、Nadav Cohen、Wei Hu 和 Yuping Luo。《深度矩阵分解中的隐式正则化》。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e Buc、E. Fox 和 R. Garnett（编），《神经信息处理系统进展》，第 32 卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.[31] Noam Razin 和 Nadav Cohen。《深度学习中的隐式正则化或许无法用范数解释》。载于 H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin（编），《神经信息处理系统进展》，第 33 卷，第 21174–21187 页。Curran Associates, Inc.，2020年。网址：https://proceedings.neurips.cc/paper/2020/file/f21e255f89e0f258accbe4e984eef486-Paper.pdf.[32]Li Jing、Jure Zbontar 和 Yann LeCun。《隐式秩最小化自编码器》。载于 H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin（编），《神经信息处理系统进展》，第 33 卷，第 14736–14746 页。Curran Associates, Inc.，2020年。网址：https://proceedings.neurips.cc/paper/2020/file/a9078e8653368c9c291ae2f8b74012e7-Paper.pdf.[33] Alireza Mousavi-Hosseini、Sejun Park、Manuela Girotti、Ioannis Mitliagkas 和 Murat A Erdogdu。《神经网络通过随机梯度下降高效学习低维表示》。载于《第十一届国际学习表征会议》，2023年。网址：https://openreview.net/forum?id=6taykzqcPD.[34]Ilya Loshchilov 和 Frank Hutter。《解耦权重衰减正则化》载于《国际学习表征会议》，2019年。网址：https://openreview.net/forum?id=Bkg6RiCqY7.[35] Edward N. Lorenz。《确定性非周期流》。《大气科学杂志》，20(2)：130–141，1963年。doi：https://doi.org/10.1175/1520-0469(1963)020¡0130:DNF¿2.0.CO;2。网址：https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.

[36]Kazumasa A. Takeuchi、Hong-liu Yang、Francesco Ginelli、G¨unter Radons 和 Hugues Chat´e。《切空间的双曲解耦与耗散系统的有效维数》

。《物理评论E》，84:046214，2011年10月。doi:10.1103/PhysRevE.84.046214。网址 https://link.aps.org/doi/10.1103/PhysRevE.84.046214。[37] Kathleen Champion、Bethany Lusch、Nathan Kutz 和 Steven L. Brunton。利用几何构造确定相空间重构的嵌入维数。《美国国家科学院院刊》，116:22445–22451，2019年10月。doi: 10.1073/pnas.1906995116。网址https://www.pnas.org/doi/10.1073/pnas.1906995116。[38] IvicaKi˘ci´c、Pantelis R. Vlachas、Georgios Arampatzis、Michail Chatzimanolakis、Leonidas Guibas 和 Petros Koumoutsakos。复杂系统有效动力学的自适应学习与在线建模。《计算机在应用力学与工程中的方法》，415:116204，2023年。ISSN 0045-7825。doi:10.1016/j.cma.2023.116204。[39] Sanjeeb T. Bose 和 George Ilhwan Park。复杂湍流的壁面模型大涡模拟。《流体力学年度评论》，50(1):535–561，2018年。ISSN 0066-4189。doi:10.1146/annurev-fluid-122316-045241。[40] Tian Qi Chen、Yulia Rubanova、Jesse Bettencourt 和 David Duvenaud。神经常微分方程。载于 Samy Bengio、Hanna M. Wallach、Hugo Larochelle、Kristen Grauman、Nicol`o Cesa-Bianchi 和 Roman Garnett（编），《神经信息处理系统进展》第31卷：2018年神经信息处理系统年度会议，NeurIPS 2018，2018年12月3日至8日，蒙特 ´利尔，加拿大，第6572–6583页，2018年。网址 https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-

摘要.htm

[41] Charles D. Young、Patrick T. Corona、Anukta Datta、Matthew E. Helgeson 和 Michael D. Graham。《拉格朗日演化过程中的散射信息驱动微观结构预测（SIM-PLE）——用于流动中复杂流体建模的数据驱动框架》。《流变学学报》，第 1–18 页，2023年。国际标准书号 0035-4511。doi:10.1007/s00397-023-01412-0。[42]Michael D Graham 和 IG Kevrekidis。用于模型降阶与数据分析的 Karhunen-Loeve分解替代方法。《计算机与化学工程》，20(5):495 506，1996年 01 月。网址：http://apps.isiknowledge.com/InboundService.do?product=WOS&action=retrieve&SrcA

pp=Papers&U

[43] E. J. Hinch。《摄动方法》。剑桥应用数学丛书。剑桥大学出版社，1991年。国际标准书号 0-521-37897-4。